ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

2025 (24) 4

қазан- желтоқсан

ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Меншік иесі: АҚ «Халықаралық ақпараттық технологиялар университеті» (Алматы қ.).
Қазақстан Республикасы Ақпарат және қоғамдық даму министрлігіне мерзімді баспасөз басылымын есепке қою туралы куәлік № KZ82VPY00020475, 20.02.2020 ж. берілген
Тақырып бағыты: ақпараттық технологиялар, ақпараттық қауіпсіздік және коммуникациялық технологиялар, әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технология.
Мерзімділігі: жылына 4 рет.
Тираж: 100 дана.
Редакция мекенжайы: 050040 Алматы қ., Манас қ., 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz
Журнал сайты: https://journal.iitu.edu.kz

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан **№ KZ82VPY00020475,** выданное от **20.02.2020 г.**
Тематическая направленность: информационные технологии, информационная безопасность и коммуникационные технологии, цифровые технологии в развитии социо-экономических систем**.**
Периодичность: 4 раза в год.
Тираж: 100 экземпляров.
Адрес редакции: 050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz Сайт журнала: https://journal.iitu.edu.kz

«International Journal of Information and Communication Technologies»
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Owner: International Information Technology University JSC (Almaty).
The certificate of registration of a periodical printed publication in the Ministry of Information and Social Development of the Republic of Kazakhstan, Information Committee No. KZ82VPY00020475, issued on 20.02.2020.
Thematic focus: information technology, digital technologies in the development of socio-economic systems, information security and communication technologies
Periodicity: 4 times a year.
Circulation: 100 copies.
Editorial address: 050040. Manas st. 34/1, Almaty. +7 (727) 244-51-09. E-mail: ijict@iitu.edu.kz
Journal website: https://journal.iitu.edu.kz

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 01–19
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.001

SECURING SOCS: UNDERSTANDING VULNERABILITIES, THEIR IMPACT, AND MITIGATION STRATEGIES

A.A. Altynbekov[1], G. Alin[2], S. Amanzholova[1], M. Saleh[1]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Astana IT University, Astana, Kazakhstan.
E-mail: 41378@iitu.edu.kz

Altynbekov Ali — master's student, Faculty of Computer Technology and Cybersecurity, International Information Technology University, Almaty,

Kazakhstan

E-mail: 41378@iitu.edu.kz. https://orcid.org/0009-0001-5360-0128;
Alin Galymzada — Candidate of Technical Sciences, assistant professor, Cybersecurity Department, International Information Technology University, Almaty, Kazakhstan
E-mail: g.alin@iitu.edu.kz. https://orcid.org/0000-0003-1028-5452;
Amanzholova Saule — Candidate of Technical Sciences, associate professor, Cybersecurity Department, Astana IT University
E-mail: s.amanzholova@astanait.edu.kz. https://orcid.org/0000-0002-6779-9393;
Saleh Mohammed — PhD, associate professor, Cybersecurity Department, International Information Technology University, Almaty, Kazakhstan
E-mail: m.saleh@iitu.edu.kz. https://orcid.org/0000-0003-4673-5056.

Abstract. Security Operations Centers (SOCs) are critical for defending orga-nizations against increasingly sophisticated cyber threats; however, they themselves are susceptible to vulnerabilities that can compromise their effectiveness. This review paper provides a comprehensive analysis of key vulnerabilities within SOC environ-ments, assesses their impact on detection and response capabilities, and explores ef-fective mitigation strategies to enhance SOC resilience. By systematically reviewing existing literature, industry reports, and case studies, the paper examines both tech-nical and organizational vulnerabilities affecting SOC performance. Common issues identified include under-resourced teams, tool misconfigurations, inefficient incident response processes, staffing shortages, and outdated technologies. The impact of these vulnerabilities is discussed in terms of delayed threat detection, increased risk of security breaches, and overall degradation of organizational cybersecurity posture. The review also synthesizes recommended mitigation strategies such as improving SOC staffing

levels, enhancing tool integration, and adopting automation technologies for incident response. Additionally, it explores the potential of advanced technologies like artificial intelligence and machine learning to enhance SOC operations and adapt to the evolving cyber threat landscape. Emphasis is placed on fostering robust com-munication protocols, promoting continuous analyst training, and integrating holistic security practices across various organizational layers. In conclusion, understanding and addressing vulnerabilities within SOCs is crucial for maintaining effective cy-bersecurity defenses. The paper underscores the need for coordinated efforts to inte-grate human expertise with technological solutions, highlighting how such synergy fortifies both prevention and response to emerging threats. Future research directions include further exploration of AI and machine learning applications in SOCs to better respond to an ever-evolving threat landscape.

Keywords: soc, vulnerabilities, threat detection, automation, artificial intelli-gence (AI), machine learning (ML), resilience

For citation: A. Altynbekov, G. Alin, S. Amanzholova, M. Saleh. Securing socs: understanding vulnerabilities, their impact, and mitigation strategies//International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 01–19 (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.001

SOC ҚАУІПСІЗДІГІ: ОСАЛДЫҚТАРДЫ, ОЛАРДЫҢ ӘСЕРІН ЖӘНЕ САЛДАРЫН АЗАЙТУ СТРАТЕГИЯЛАРЫН ТҮСІНУ

А.А. Алтынбеков[1], Г. Алин[2], С. Аманжолова[1], М. Салех[1]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]Астана IT Университеті, Астана, Қазақстан.
E-mail: 41378@iitu.edu.kz

Алтынбеков Али — магистрант, Компьютерлік технологиялар және киберқауіпсіздік факультеті, Ақпараттық технологиялар халықаралық университеті
E-mail: 41378@iitu.edu.kz. https://orcid.org/0009-0001-5360-0128;
Алин Галымзада — техника ғылымдарының кандидаты, киберқауіпсіздік кафедрасының ассистент профессоры, Халықаралық ақпараттық технологиялар университеті
E-mail: g.alin@iitu.edu.kz. https://orcid.org/0000-0003-1028-5452;
Аманжолова Сауле — техника ғылымдарының кандидаты, Киберқауіпсіздік кафедрасының қауымдастырылған профессоры, Астана IT университеті
E-mail: s.amanzholova@astanait.edu.kz. https://orcid.org/0000-0002-6779-9393;
Салех Мохаммед — PhD, Киберқауіпсіздік кафедрасының қауымдастырылған профессоры, Ақпараттық технологиялар халықаралық университеті
E-mail: m.saleh@iitu.edu.kz. https://orcid.org/0000-0003-4673-5056.

Аннотация. Қауіпсіздік Операциялары Орталықтары (SOCs) ұйымдар-ды барған сайын күрделі киберқауіптерден қорғау үшін өте маңызды; дегенмен, олардың өздері олардың тиімділігіне нұқсан келтіруі мүмкін осалдықтарға бей-ім. Бұл шолу мақаласы SOC орталарындағы негізгі осалдықтардың жан-жақты талдауын қамтамасыз етеді, олардың анықтау және әрекет ету мүмкіндіктері-не әсерін бағалайды және SOC тұрақтылығын арттыру үшін тиімді жұмсар-ту стратегияларын зерттейді. Қолданыстағы әдебиеттерді, салалық есептерді және кейстерді жүйелі түрде қарастыра отырып, мақалада SOC өнімділігіне әсер ететін техникалық және ұйымдастырушылық осалдықтар қарастырылады. Анықталған жалпы мәселелерге ресурстардың жетіспеушілігі, құралдардың дұрыс конфигурацияланбауы, инциденттерге әрекет етудің тиімсіз процестері, қызметкерлердің жетіспеушілігі және ескірген технологиялар жатады. Бұл осал-дықтардың әсері қауіптерді анықтаудың кешігуі, қауіпсіздіктің бұзылу қаупінің жоғарылауы және ұйымның киберқауіпсіздік жағдайының жалпы нашарлауы тұрғысынан талқыланады. Шолу СОНЫМЕН қатар SOC қызметкерлерінің са-нын жақсарту, құралдардың интеграциясын жақсарту және инциденттерге жау-ап беру үшін автоматтандыру технологияларын енгізу сияқты ұсынылған жұм-сарту стратегияларын синтездейді. Сонымен қатар, ол SOC жұмысын жақсарту және дамып келе жатқан киберқауіптер ландшафтына бейімделу үшін жасан-ды интеллект және машиналық оқыту сияқты озық технологиялардың әлеуетін зерттейді. Сенімді байланыс хаттамаларын әзірлеуге, талдаушыларды үздіксіз оқытуға жәрдемдесуге және әртүрлі ұйымдық деңгейлерде қауіпсіздіктің тұтас әдістерін біріктіруге баса назар аударылады. Қорытындылай келе, Soc жүйесін-дегі осалдықтарды түсіну және жою киберқауіпсіздікті қорғаудың тиімді құрал-дарын сақтау үшін өте маңызды. Құжатта адам тәжірибесін технологиялық шешімдермен біріктіру бойынша үйлестірілген күш-жігердің қажеттілігі атап өтіліп, мұндай синергия пайда болатын қауіптердің алдын алуды да, оларға жа-уап беруді де қалай күшейтетіні көрсетілген. Болашақ зерттеу бағыттары үнемі өзгеріп отыратын қауіп-қатер ландшафтына жақсырақ жауап беру үшін Soc-да жасанды интеллект пен машиналық оқыту қолданбаларын одан әрі зерт-теуді қамтиды.

Түйін сөздер: soc, осалдықтар, қауіптерді анықтау, автоматтандыру, жасанды интеллект, машиналық оқыту, төзімділік

Дәйексөздер үшін: А. Алтынбеков, Г. Алин, С. Аманжолова, М. Салех. SOC қауіпсіздігі: осалдықтарды, олардың әсерін және салдарын азайту стратегияларын түсіну // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Т. 6. № 24. Б. 01–19 (ағыл. тілінде). https://doi.org/10.54309/IJICT.2025.24.4.001

ЗАЩИТА SOCS: ПОНИМАНИЕ УЯЗВИМОСТЕЙ, ИХ ВОЗДЕЙСТВИЯ И СТРАТЕГИЙ СМЯГЧЕНИЯ ПОСЛЕДСТВИЙ

А.А. Алтынбеков[1], Г. Алин[2], С. Аманжолова[1], М. Салех[1]
[1]Международный университет информационных технологий, Алматы, Казахстан;
[2]Астана ИТ университет, Астана, Казахстан.
E-mail: 41378@iitu.edu.kz

Алтынбеков Али — магистратрант, факультет компьютерных технологий и кибербезопасности, Международный университет информационных технологий, Алматы, Казахстан
E-mail: 41378@iitu.edu.kz. https://orcid.org/0009-0001-5360-0128;
Алин Галымзада — кандидат технических наук, ассистент профессор, кафедра кибербезопасности, Международный университет информационных технологий, Алматы, Казахстан
E-mail: g.alin@iitu.edu.kz. https://orcid.org/0000-0003-1028-5452;
Аманжолова Сауле — кандидат технических наук, ассоциированный профессор, кафедра кибербезопасности, Астана ИТ университет, Астана, Казахстан
E-mail: s.amanzholova@astanait.edu.kz. https://orcid.org/0000-0002-6779-9393;
Салех Мохаммед — PhD, ассоциированный профессор, кафедра кибербезопасности, Международный университет информационных технологий,
E-mail: m.saleh@iitu.edu.kz. https://orcid.org/0000-0003-4673-5056.

Аннотация. Операционные центры обеспечения устойчивости (SOCs) имеют решающее значение для защиты организаций от все более изощренных киберугроз; однако они сами подвержены уязвимостям, которые могут поста-вить под угрозу их эффективность. В этом обзорном документе представлен все-сторонний анализ ключевых уязвимостей в среде SOC, оценивается их влияние на возможности обнаружения и реагирования, а также рассматриваются эффек-тивные стратегии смягчения последствий для повышения устойчивости SOC. На основе систематического анализа существующей литературы, отраслевых отчетов и тематических исследований в документе рассматриваются как технические, так и организационные уязвимости, влияющие на производительность SOC. Выявленные общие проблемы включают нехватку ресурсов у команд, не-правильную настройку инструментов, неэффективные процессы реагирования на инциденты, нехватку персонала и устаревшие технологии. Влияние этих уязвимостей обсуждается с точки зрения задержки обнаружения угроз, повы-шенного риска нарушений безопасности и общего ухудшения состояния

кибербезопасности организации. В обзоре также обобщены рекомендуемые страте-гии смягчения последствий, такие как повышение уровня укомплектованности SOC персоналом, усиление интеграции инструментов и внедрение технологий автоматизации для реагирования на инциденты. Кроме того, в обзоре рассматривается потенциал передовых технологий, таких как искусственный интеллект и машинное обучение, для улучшения функционирования SOC и адаптации к меняющемуся ландшафту киберугроз. Особое внимание уделяется созданию надежных коммуникационных протоколов, непрерывному обучению аналитиков и интеграции целостных методов обеспечения безопасности на различных уровнях организации. В заключение, понимание и устранение уязвимостей в SOCs имеет решающее значение для поддержания эффективной защиты от кибербезопасности. В документе подчеркивается необходимость скоординированных усилий по интеграции человеческого опыта с технологиче-скими решениями, подчеркивая, как такая синергия способствует предотвраще-нию и реагированию на возникающие угрозы. Будущие направления исследова-ний включают дальнейшее изучение приложений искусственного интеллекта и машинного обучения в социальных сетях, чтобы лучше реагировать на постоянно меняющийся ландшафт угроз.

Ключевые слова: soc, уязвимости, обнаружение угроз, автоматизация, искусственный интеллект, машинное обучение, устойчивость

Для цитирования: А. Алтынбеков, Г. Алин, С. Аманжолова, М. Салех. Защита SOCS: понимание уязвимостей, их воздействия и стратегий смягчения последствий // Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 01–19. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.001

Introduction

In today's digital landscape, where technological innovation is a key driv-er of organizational success, the proliferation of cyber threats presents significant challenges to maintaining secure information systems (Riggs et al., 2023; Aslan & Samet, 2017). Cyber adversaries are increasingly sophisticated, employing advanced techniques such as zero-day exploits, ransomware, and state-sponsored attacks to ex-ploit vulnerabilities within organizational infrastructures (Grobler et al., 2018; De-mertzis et al., 2018). To defend against these evolving threats, organizations deploy Security Operations Centers (SOCs), which serve as centralized units for continuous monitoring, detection, analysis, and response to cybersecurity incidents (Agyepong et al., 2020; Muniz et al., 2015). SOCs combine skilled personnel, cutting-edge technologies, and structured processes to provide comprehensive cybersecurity defense mechanisms (Onwubiko, 2015; Vielberth et al., 2020).

Despite their critical role, SOCs face inherent vulnerabilities that can diminish their operational effectiveness (Kokulu et al., 2019; Janos & Dai, 2018).

Technical challenges such as tool misconfigurations, lack of interoperability between security solutions, and reliance on legacy systems can impede the SOC's ability to detect and respond to threats promptly (Banati et al., 2022; Kiselev et al., 2022). For instance, Banati et al. discuss how the use of attack graphs can optimize SOC operations by ad-dressing technical inefficiencies (Banati et al., 2022). Organizational issues, including insufficient staffing levels, high turnover rates, inadequate training, and lack of clear communication channels, further exacerbate these technical shortcomings (Reeves & Ashenden, 2023; Onwubiko & Ouazzane, 2019). Kokulu et al. highlight that mis-matched SOCs, where the organizational structure does not align with operational needs, are particularly prone to such vulnerabilities (Kokulu et al., 2019).

The consequences of these vulnerabilities are substantial. Delays in threat detection and response due to internal SOC weaknesses can result in significant financial losses, operational disruptions, regulatory penalties, and erosion of stakeholder trust (Anton et al., n.d.; Resilience, 2024). As cyber threats become more advanced and persistent, the inability of SOCs to adapt and address internal vulnerabilities pos-es a critical risk to organizational security (Vielberth et al., 2020; Onwubiko, 2017). Despite the recognition of these challenges, there is a scarcity of comprehensive stud-ies that systematically analyze SOC vulnerabilities and their impact on cybersecurity operations (Vielberth et al., 2020; Taqafi et al., 2023). Addressing this research gap is imperative for developing effective strategies to enhance SOC resilience and efficacy (Aslan & Samet, 2017; Hore et al., 2023).

This review paper, titled "Securing SOCs: Understanding Vulnerabilities, Their Impact, and Mitigation Strategies," aims to bridge this gap by providing a thor-ough examination of the vulnerabilities affecting SOCs and proposing actionable solutions. The primary objectives are to identify and categorize the key technical and organizational vulnerabilities within SOC environments, assess the impact of these vulnerabilities on SOC performance, and explore effective mitigation strategies that can enhance SOC capabilities and resilience (Vielberth et al., 2020; Taqafi et al., 2023). By synthesizing current research and best practices, this paper seeks to con-tribute to the advancement of SOC effectiveness in the face of evolving cyber threats.

To achieve these objectives, the paper seeks to answer several research questions: What are the most prevalent technical and organizational vulnerabilities affect-ing SOCs today? How do these vulnerabilities impact the operational effectiveness of SOCs in detecting and responding to cyber threats? What strategies can be imple-mented to mitigate these vulnerabilities and improve SOC performance? How can emerging technologies such as artificial intelligence (AI) and machine learning (ML) be leveraged to address SOC vulnerabilities?

To address the research objectives systematically, the paper is organized into

three main sections. Section 2 will cover the methodology of literature collection of papers, detailing the criteria for selecting relevant studies and the approaches used to gather and analyze data. Section 3 should include the results of the paper's analysis, presenting an in-depth discussion of the identified technical and organizational vulnerabilities affecting SOCs, their impact on detection and response capabilities, and the mitigation strategies proposed in the literature. The last section will be the conclusion, summarizing the key findings of the study, discussing implications for cybersecurity professionals, and suggesting future research directions, particularly in exploring the role of AI and ML in enhancing SOC operations.

Methodology

This section outlines the systematic approach to collecting and analyzing lit-erature on vulnerabilities within Security Operations Centers (SOCs) and their mit-igation strategies. The methodology was designed to provide a comprehensive and unbiased review of research, industry reports, and case studies.

The literature search strategy employed multiple reputable platforms. Google Scholar served as a broad resource for scholarly works across disciplines. IEEE Xplore provided access to technical literature in computer science and engineering, critical for sourcing papers on SOCs. Scopus covered peer-reviewed journals, conference proceedings, and books, ensuring diverse and high-quality sources. Advanced AI tools were also utilized to aggregate and summarize recent publications, capturing the latest developments in cybersecurity.

To ensure the relevance and quality of selected literature, inclusion criteria focused on publications from 2010 onward, emphasizing peer-reviewed articles, industry reports, and case studies on SOC vulnerabilities and mitigation strategies, including the use of AI and machine learning. Exclusion criteria eliminated non-English articles and publications lacking substantial contributions or relevance to SOCs. Keywords and phrases were grouped into categories—SOC vulnerabilities, impact analysis, and mitigation strategies—with Boolean operators enhancing search efficiency.

The initial search yielded numerous publications, with duplicates removed using reference management tools. A two-stage screening process followed: first, articles were assessed by titles and abstracts for relevance; then, selected articles underwent fulltext review to confirm suitability. Extracted data were organized into thematic categories such as technical vulnerabilities (e.g., outdated technologies, tool misconfigurations), organizational challenges (e.g., staffing shortages), impacts on SOC performance, and mitigation strategies like automation and AI applications. As shown in figure 1.

Data extraction involved identifying critical information such as specific tech-nical and organizational challenges and their direct effects on SOC performance met-rics like threat detection and incident response times. This process also highlighted the various technological and operational strategies employed by organizations to mitigate these vulnerabilities. For example, advancements in machine learning algo-rithms were often cited as a means of improving SOC efficiency through realtime threat detection and proactive incident response measures.

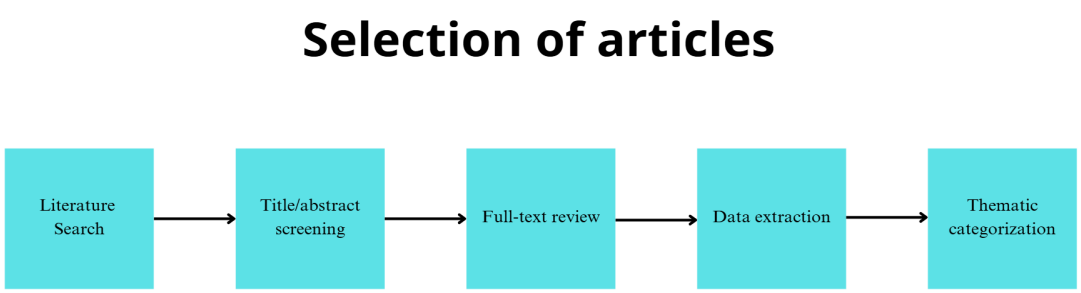


Fig. 1. Selection of articles

These findings were com-pared across studies to evaluate their effectiveness and applicability in different orga-nizational contexts.

A qualitative synthesis identified patterns, gaps, and insights across studies, focusing on the interplay between vulnerabilities and their cumulative effects on SOC effectiveness. This approach provided a holistic understanding of challenges and pro-posed solutions to enhance resilience.

Certain limitations were acknowledged. The reliance on published literature may exclude unpublished or organizational insights, introducing potential publication bias. Restricting the search to English-language studies could omit valuable interna-tional research. Additionally, the rapid evolution of cybersecurity means findings may require ongoing updates.

Ethical considerations guided the research process. Proper citation ensured acknowledgment of original authors, maintaining academic integrity. The study ad-hered to ethical guidelines, emphasizing transparency and respect for intellectual property.

This methodology offers a thorough analysis of SOC vulnerabilities and miti-gation strategies, contributing valuable insights to help organizations strengthen their defenses in an evolving cybersecurity landscape.

Results

Literature Review. Securing Security Operations Centers (SOCs) involves understanding their vulnerabilities, the impact of these vulnerabilities, and implementing effective mitigation strategies. SOCs are integral to an organization's cybersecurity strategy, providing real-time threat detection and response capabilities. However, they face challenges such as technological constraints, procedural complexities, and human factors, which can hinder their effectiveness. Addressing these challenges requires a comprehensive approach that includes optimizing processes, leveraging technology, and enhancing human expertise.

Various studies have extensively explored the development, optimization, and challenges associated with Security Operations Centers (SOCs) to enhance cybersecurity resilience. Agyepong et al. provide a comprehensive overview of SOC concepts and implementation strategies, emphasizing their critical role in organizational security frameworks (Agyepong et al., 2020). Similarly, Muniz, McIntyre, and AlFardan delve into the practical aspects of building, operating, and maintaining SOCs, offering valuable insights for practitioners aiming to establish robust security infrastructures (Muniz et al., 2015).

Addressing vulnerabilities is a focal point in cybersecurity research. Anton et al. introduce the Vulnerability Assessment and Mitigation Methodology, highlighting systematic approaches to identify and mitigate security risks effectively (Anton et al., n.d.). Aslan and Samet focus on mitigating cyber attacks by raising awareness of vulnerabilities and bugs, suggesting proactive defense mechanisms to strengthen security postures (Aslan & Samet, 2017). Degrass presents innovative techniques for identifying and remediating operational vulnerabilities, contributing to proactive security measures within organizations (E Degrass - US Patent App. 17/430 & 2022, 2022). Hore et al. propose methods for optimal triage and mitigation of context-sensitive cyber vulnerabilities, aiming to prioritize threats and allocate resources efficiently (Hore et al., 2023).

For SOC optimization, Banati et al. examine the use of attack graphs to enhance SOC operations, demonstrating how visualization tools can aid in threat detection and response (Banati et al., 2022). Demertzis et al. discuss the next-generation cognitive SOC, utilizing network flow forensics and cybersecurity intelligence to improve response times and decision-making processes (Demertzis et al., 2018). Li and Hsieh explore the implementation of distributed hierarchical SOCs using mobile agent groups, offering solutions for scalable and flexible security management (Li & Hsieh, 2010). Yuan and Zou highlight the role of correlation analysis in SOCs, showing how data integration can improve security operations and incident detection rates (Yuan & Zou, 2011).

Operational challenges within SOCs are further explored by Kiselev, Korotkikh, and Shott, who share practical experiences in establishing a SOC, discussing best practices and common pitfalls (Kiselev et al., 2022). Kokulu et al. conduct a qualitative study on SOC issues, identifying mismatches and challenges that can hinder security efforts and proposing solutions to address them (Kokulu et al., 2019). Janos and Dai investigate security concerns related to SOCs, providing insights into potential vulnerabilities and areas needing improvement (Janos & Dai, 2018).

Understanding human factors in SOCs, Reeves and Ashenden investigate decision-making processes within these centers, advocating for the use of cyber deception technology to enhance threat detection and response strategies (Reeves & Ashenden, 2023). Taqafi, Maleh, and Ouazzane propose a maturity capability framework for SOCs, aiming to assess and improve their effectiveness systematically (Taqafi et al., 2023). Vielberth et al. conduct a systematic study on SOCs, identifying open challenges and areas for future research, emphasizing the need for continuous development in SOC practices (Vielberth et al., 2020).

In the context of critical infrastructure protection, Riggs et al. (Riggs et al., 2023) discuss the impact of vulnerabilities and present mitigation strategies to ensure cyber-secure environments, highlighting the importance of SOCs in safeguarding essential services. Grobler, Jacobs, and van Niekerk emphasize the role of cybersecurity centers in threat detection and mitigation, providing insights into their implementation and operational effectiveness (Grobler et al., 2018). Resilience focuses on opti-

mizing SOCs for enhanced cyber resilience, discussing strategies for improvement and adaptation in the face of evolving threats (Resilience, 2024).

Onwubiko underscores the importance of security monitoring, situational awareness, and threat intelligence within SOCs to support cyber defense strategies (Onwubiko, 2015; Onwubiko, 2017). Together with Ouazzane, Onwubiko identifies challenges in building effective SOCs, highlighting organizational structure, resource allocation, and the integration of advanced technologies as key factors (Onwubiko & Ouazzane, 2019).

Schönfeld discusses the role of SOCs in the medical technology field, highlighting their importance in securing sensitive data and complying with regulatory requirements (Schönfeld, 2023). This sector-specific insight demonstrates the broad applicability of SOC principles across different industries.

Collectively, this body of work underscores the evolving nature of SOCs and the continuous efforts by researchers and practitioners to address cybersecurity threats effectively. The studies highlight the importance of proactive vulnerability management, advanced threat detection techniques, and the integration of human factors in enhancing the effectiveness of SOCs. As cybersecurity threats become more sophisticated, the insights from these studies provide a foundation for developing resilient and adaptive security operations capable of protecting critical assets and information. All the articles studied above can be seen in a simplified form on table 1.

Table 1. Summary of Key Contributions and Relevance of Reviewed SOC Literature

Source/ Author(s)	Key Contribution	Relevance to SOC Challenges
Agyepong et al.	Overview of SOC concepts and implementation strategies	Critical role in organizational security frameworks
Muniz, McIntyre, and AlFardan	Practical aspects of building, operating, and maintaining SOCs	Insights for robust security infrastructures
Anton et al.	Vulnerability Assessment and Mitigation Methodology	Systematic approach to identify and mitigate risks
Aslan and Samet	Proactive defense mechanisms for mitigating vulnerabilities	Strengthens cybersecurity postures
Degrass	Innovative techniques for operational vulnerability remediation	Contributes to proactive security measures
Hore et al.	Optimal triage and mitigation of context-sensitive vulnerabilities	Efficient threat prioritization and resource allocation
Banati et al.	Use of attack graphs to enhance SOC operations	Improves threat detection and response efficiency
Demertzis et al.	Next-generation cognitive SOC using network flow forensics	Enhances decision-making and response times
Li and Hsieh	Distributed hierarchical SOCs using mobile agent groups	Solutions for scalable and flexible SOC management

Yuan and Zou	Role of correlation analysis in improving SOC performance	Improves incident detection rates through data integration
Kiselev, Korotkikh, and Shott	Practical experiences in establishing SOCs	Best practices and common pitfalls in SOC implementation
Kokulu et al.	Qualitative study on SOC challenges and solutions	Addresses mismatches and challenges in SOC operations
Janos and Dai	Insights into security concerns related to SOCs	Highlights vulnerabilities and improvement areas
Reeves and Ashenden	Decision-making processes in SOCs with cyber deception technology	Enhances threat detection and response strategies
Taqafi, Maleh, and Ouazzane	Maturity capability framework for assessing SOC effectiveness	Systematically assesses and improves SOC effectiveness
Vielberth et al.	Systematic study identifying open challenges in SOCs	Emphasizes need for continuous SOC development
Riggs et al.	Mitigation strategies for vulnerabilities in critical infrastructure	Focus on safeguarding critical services
Grobler, Jacobs, and van Niekerk	Role of cybersecurity centers in threat detection and mitigation	Insights into implementation and operational effectiveness
Resilience	Strategies for enhancing SOC resilience	Adaptation strategies for evolving threats
Onwubiko (2015), (2017)	Importance of situational awareness and threat intelligence	Supports effective monitoring and cyber defense
Ouazzane and Onwubiko	Challenges in building effective SOCs	Focus on structure, resources, and advanced technologies
Schönfeld	Role of SOCs in medical technology security and compliance	Broad applicability across industries

The systematic review of the literature revealed a comprehensive understanding of the vulnerabilities affecting Security Operations Centers (SOCs), their impact on organizational cybersecurity, and the strategies available for mitigation. The findings are organized into three main categories as shown on Figure 2: technical vulnerabilities, organizational vulnerabilities, and mitigation strategies, including the potential role of advanced technologies such as artificial intelligence (AI) and machine learning (ML).

Technical Vulnerabilities in SOCs

One of the most prevalent technical vulnerabilities identified is the misconfiguration of security tools within SOCs. Misconfigurations can lead to blind spots in security monitoring, allowing threats to go undetected (Banati et al., 2022; Kiselev et al., 2022). Banati et al. highlight that improper setup of intrusion detection systems and firewalls can result in false negatives, where malicious activities are not flagged, compromising the organization's security posture (Banati et al., 2022).

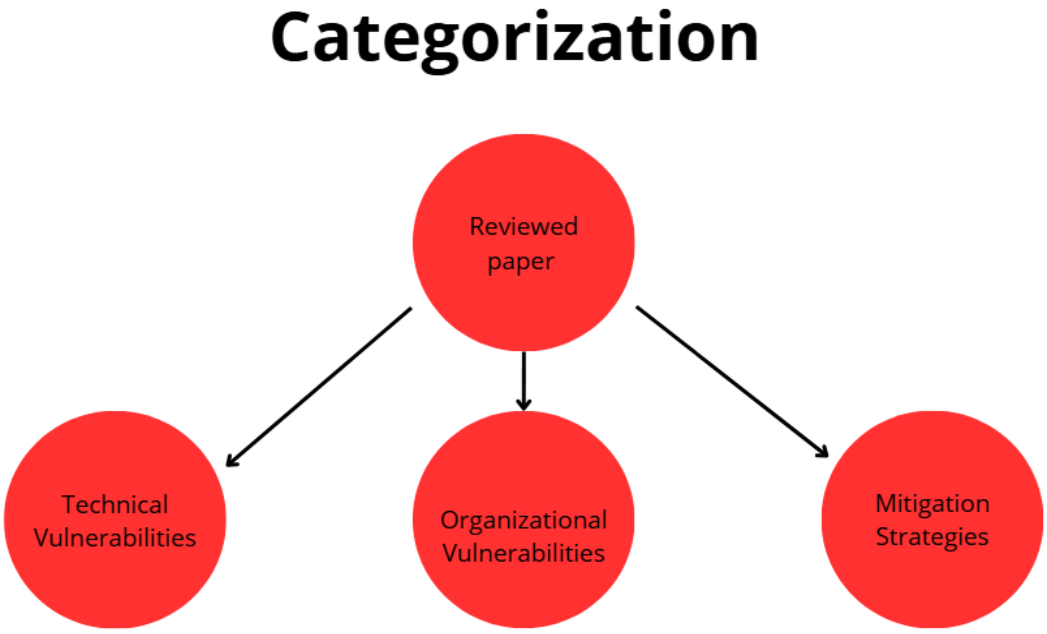


Fig. 2. Categorization of revied papers

The integration of disparate security tools poses a significant challenge. Many SOCs rely on multiple security solutions that do not seamlessly communicate with each other, leading to information silos and inefficient incident response processes (Demertzis et al., 2018; Yuan & Zou, 2011). Demertzis et al. emphasize that the lack of interoperability hinders the SOC's ability to correlate events across different platforms, reducing the effectiveness of threat detection and analysis (Demertzis et al., 2018).

Dependence on outdated technologies and legacy systems is another critical vulnerability. Legacy systems often lack the necessary features to combat modern cyber threats and may not receive regular security updates, making them susceptible to exploits (Kiselev et al., 2022; Li & Hsieh, 2010). Kiselev et al. note that legacy infrastructure can impede the adoption of new security solutions, limiting the SOC's capability to respond to advanced persistent threats (Kiselev et al., 2022).

Organizational Vulnerabilities in SOCs

Under-resourced teams are a common organizational vulnerability within SOCs. Insufficient staffing leads to overworked analysts, increased fatigue, and a higher likelihood of oversight in monitoring activities (Kokulu et al., 2019), (Onwubiko & Ouazzane, 2019). Kokulu et al. found that staffing shortages result in delayed incident responses and reduced overall efficiency in security operations (Kokulu et al., 2019).

The cybersecurity field faces high turnover rates due to factors such as burnout, competitive job markets, and inadequate career development opportunities. High turnover disrupts team cohesion and results in the loss of institutional knowledge,

weakening the SOC's effectiveness (Janos & Dai, 2018; Reeves & Ashenden, 2023). Reeves and Ashenden suggest that retaining skilled personnel is crucial for maintaining robust security operations (Reeves & Ashenden, 2023).

A lack of continuous training and professional development leaves SOC personnel ill-equipped to handle emerging threats. Without up-to-date knowledge of the latest attack vectors and security technologies, analysts may fail to identify or respond appropriately to incidents (Agyepong et al., 2020; Onwubiko, 2017). Onwubiko stresses the importance of ongoing education to enhance situational awareness and threat intelligence capabilities (Onwubiko, 2017).

Effective communication is essential for coordinated incident response. The absence of clear communication protocols within the SOC and between departments can lead to confusion and delayed actions during critical incidents (Kokulu et al., 2019; Onwubiko & Ouazzane, 2019). Kokulu et al. highlight that organizational silos hinder information sharing, which is vital for a timely and effective security response (Kokulu et al., 2019).

Impact of SOC Vulnerabilities on Detection and Response Capabilities

The identified vulnerabilities have a profound impact on the SOC's ability to detect and respond to cyber threats effectively.

Technical and organizational shortcomings contribute to slower identification of security incidents. Delayed threat detection allows attackers more time to infiltrate systems, exfiltrate data, and cause damage (Anton et al., n.d.; Resilience, 2024).

Figure 3 illustrates the average detection times across industries based on findings by Anton et al. and Resilience (Anton et al., n.d.; Resilience, 2024). For example, the healthcare industry reports the longest detection time, averaging 211 days. Such delays highlight critical vulnerabilities in identifying security breaches promptly.

Anton et al. indicate that timely detection is critical in minimizing the impact of security breaches (Anton et al., n.d.).

Vulnerabilities within SOCs heighten the risk of successful cyber attacks. Inefficient processes and inadequate defenses provide opportunities for adversaries to exploit weaknesses, leading to potential financial losses and reputational damage (Grobler et al., 2018; Riggs et al., 2023). Riggs et al. emphasize that compromised SOCs can have cascading effects on critical infrastructure protection (Riggs et al., 2023).

The cost of delayed detection can be significant. For instance, E. Resilience (Resilience, 2024) reports that organizations experiencing major data breaches incur average costs of $4.45 million per incident. These expenses include legal fees, regulatory penalties, lost revenue, and the cost of remediation efforts. The financial toll reinforces the necessity of investing in advanced detection technologies and staff training.

Figure 4 presents the cost breakdown of data breaches as reported by Resilience (Resilience, 2024).

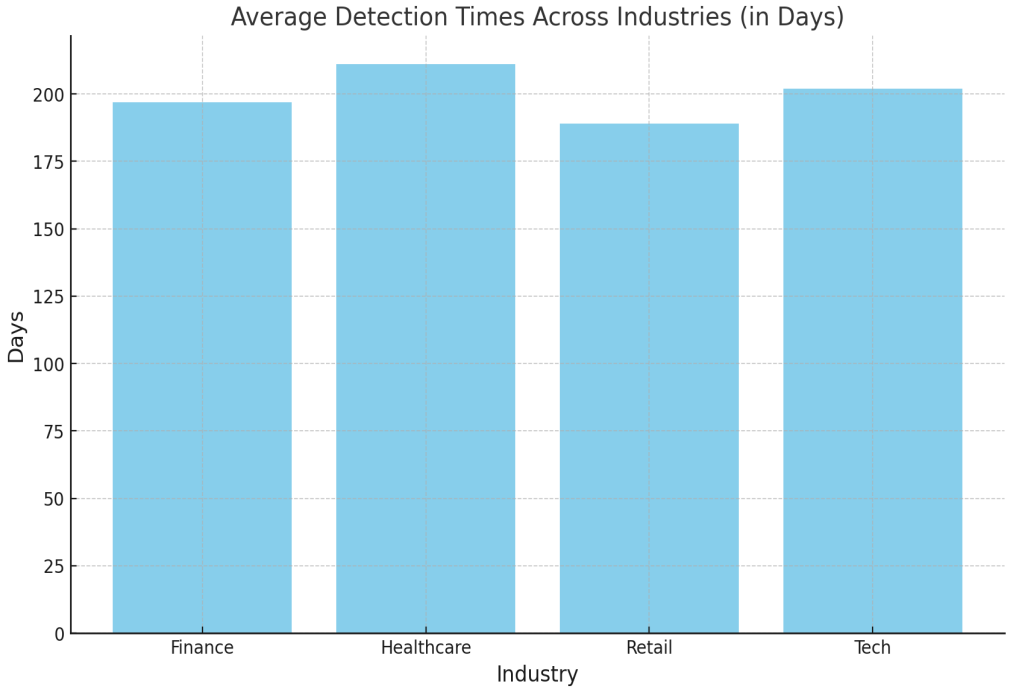


Fig. 3. Average Detection Times Across Industries

Lost revenue constitutes the largest portion (33.7 %), followed by legal fees (27.0 %). These statistics underscore the financial urgency of addressing SOC vulnerabilities.

Persistent vulnerabilities erode the overall cybersecurity posture of an organization. A weakened SOC cannot adequately protect assets, comply with regulatory requirements, or maintain stakeholder trust (Onwubiko, 2015; Vielberth et al., 2020). Vielberth et al. note that organizations with vulnerable SOCs are less resilient to cyber threats and may struggle to recover from incidents (Vielberth et al., 2020).

Mitigation Strategies for Enhancing SOC Resilience

The literature suggests several strategies to address the identified vulnerabilities and improve SOC performance.

Addressing staffing shortages involves not only hiring additional personnel but also optimizing workforce allocation. Implementing flexible staffing models and utilizing remote analysts can expand coverage and reduce burnout (Onwubiko & Ouazzane, 2019; Taqafi et al., 2023). Taqafi et al. propose a maturity capability framework to assess and enhance staffing effectiveness within SOCs (Taqafi et al., 2023).

Investing in security platforms that offer interoperability and centralized management can mitigate technical vulnerabilities related to tool misconfigurations and integration issues (Demertzis et al., 2018; Yuan & Zou, 2011). Yuan and Zou demonstrate how correlation analysis and integrated systems improve incident detection rates and response times (Yuan & Zou, 2011).

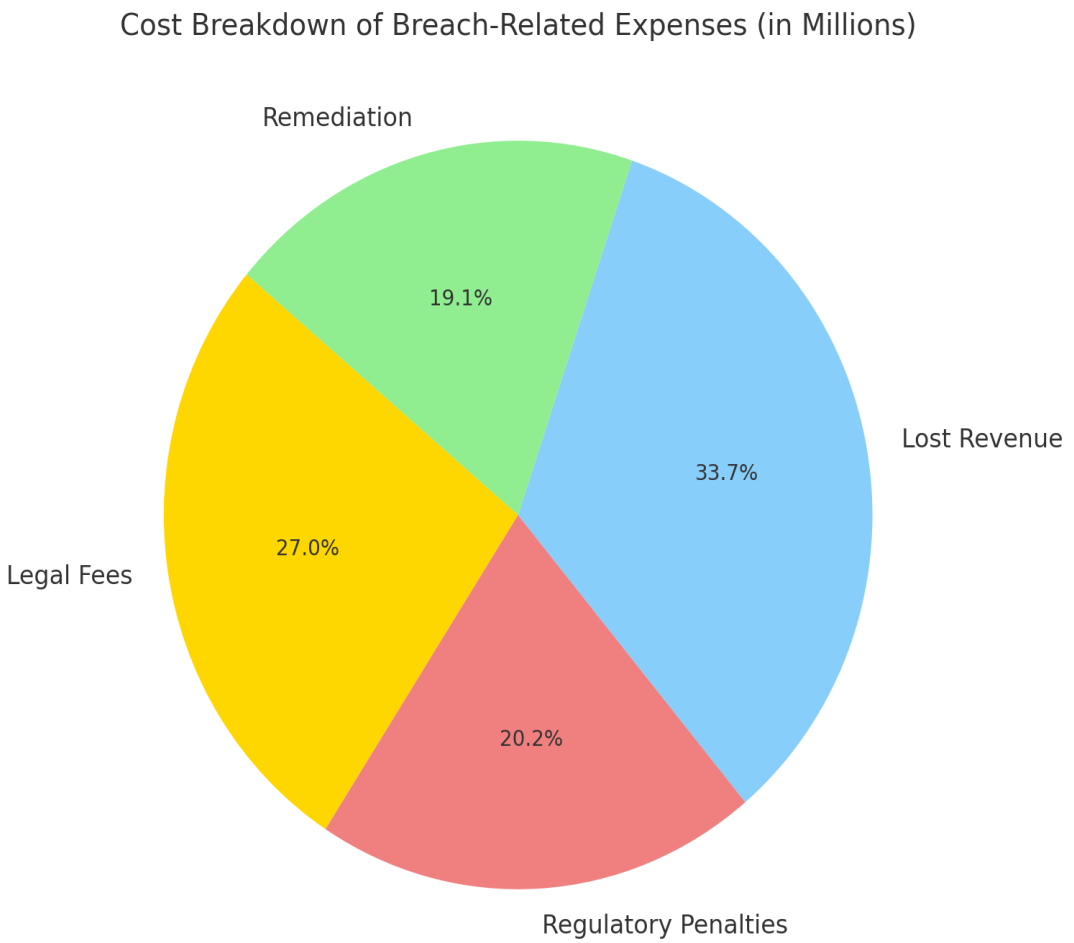


Fig. 4. Cost Breakdown of Breach-Related Expenses (in Millions)

Automation reduces the manual workload on analysts, allowing them to focus on complex threat analysis. Implementing automated incident response workflows and utilizing security orchestration tools can streamline operations and reduce the potential for human error (Banati et al., 2022; E Degrass - US Patent App. 17/430 & 2022, 2022). Degrass highlights the effectiveness of automation in proactive vulnerability remediation (E Degrass - US Patent App. 17/430 & 2022, 2022).

The application of AI and ML offers significant potential in enhancing SOC capabilities. These technologies can analyze vast amounts of data to identify patterns indicative of cyber threats, predict potential vulnerabilities, and recommend mitigation actions (Aslan & Samet, 2017; Demertzis et al., 2018). Aslan and Samet discuss how AI-driven solutions can augment human analysts, providing advanced threat detection and response mechanisms (Aslan & Samet, 2017).

Investing in continuous training programs ensures that SOC personnel remain knowledgeable about the latest threats and technologies. Certifications, workshops,

and simulation exercises can improve analysts' skills and confidence in handling incidents (Agyepong et al., 2020; Onwubiko, 2017). Onwubiko emphasizes the role of education in developing situational awareness and effective security monitoring (Onwubiko, 2015).

Developing standardized communication procedures facilitates better coordination within the SOC and across the organization. Regular meetings, incident debriefings, and collaborative platforms can enhance information sharing and response efficiency (Kokulu et al., 2019; Reeves & Ashenden, 2023). Reeves and Ashenden advocate for integrating communication strategies into SOC workflows to improve decision-making processes (Reeves & Ashenden, 2023).

Potential of Advanced Technologies in SOC Operations

The integration of advanced technologies such as AI and ML is emerging as a key strategy for enhancing SOC resilience.

Figure 5 highlights the adoption rates of advanced technologies like AI and ML, as reported by Aslan and Samet, Demertzis et al., and Hore et al. (Demertzis et al., 2018; Aslan & Samet, 2017; Hore et al., 2023). The timeline demonstrates the progression from basic vulnerability awareness in 2017 to sophisticated predictive analytics in 2023, illustrating the growing reliance on these technologies to enhance SOC operations.

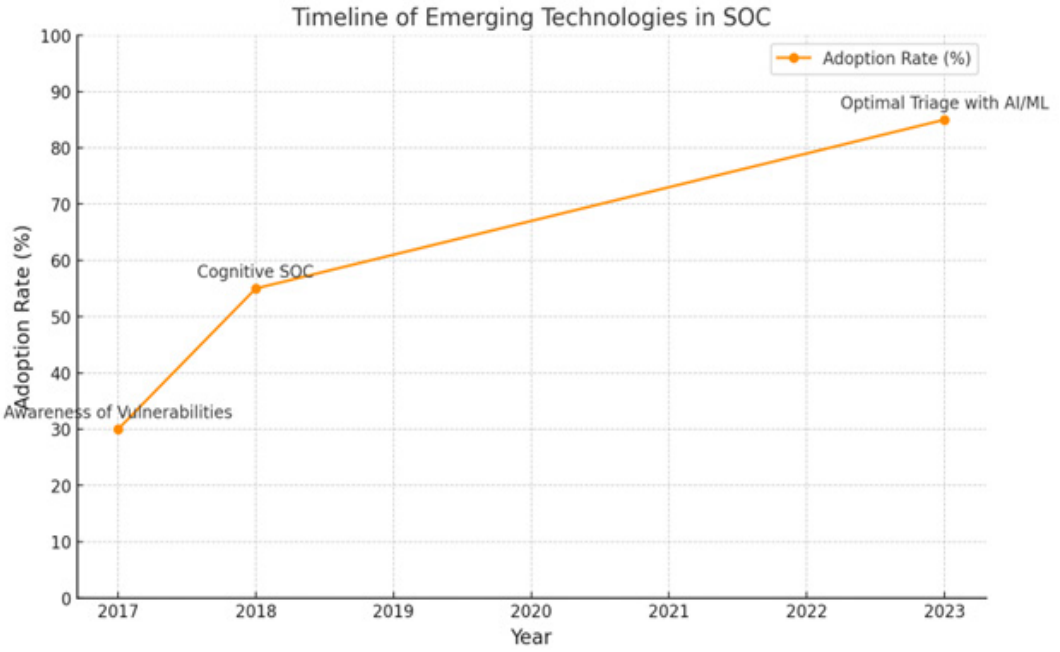


Fig. 5. Timeline of Emerging Technologies in SOC

AI and ML can process large datasets to identify trends and anomalies that may signify security threats. Predictive analytics enable SOCs to anticipate attacks and proactively strengthen defenses (Demertzis et al., 2018; Hore et al., 2023). Hore

et al. demonstrate the use of ML algorithms in optimizing vulnerability triage and mitigation efforts (Hore et al., 2023).

Machine learning models can automate routine tasks, such as alert prioritization and initial incident response actions. This automation accelerates response times and allows human analysts to concentrate on complex investigations (E Degrass - US Patent App. 17/430 & 2022, 2022; Vielberth et al., 2020). Vielberth et al. discuss the role of cognitive SOCs that leverage AI for intelligent decision-making (Vielberth et al., 2020).

AI technologies can augment human capabilities by providing insights and recommendations, effectively bridging gaps caused by staffing shortages or skill deficiencies (Aslan & Samet, 2017; Demertzis et al., 2018). Aslan and Samet highlight that AI does not replace human analysts but enhances their ability to manage threats more effectively (Aslan & Samet, 2017).

The results of this review underscore the multifaceted nature of vulnerabilities within SOCs, encompassing both technical and organizational dimensions. These vulnerabilities significantly impair the SOC's ability to detect and respond to cyber threats, necessitating a comprehensive approach to mitigation. Strategies that combine improvements in staffing, technology integration, automation, and training are essential for enhancing SOC resilience. The incorporation of AI and ML technologies presents a promising avenue for advancing SOC capabilities, offering tools to manage the increasing volume and complexity of cyber threats.

Discussion

The findings highlight the multifaceted vulnerabilities Security Operations Centers (SOCs) face, spanning both technical and organizational domains. Technical issues like tool misconfigurations, fragmented security solutions, and legacy infrastructures limit comprehensive threat detection and response capabilities, creating detection blind spots and complicating incident correlation (Demertzis et al., 2018), (Banati et al., 2022; Kiselev et al., 2022; Yuan & Zou, 2011). Meanwhile, organizational challenges—such as insufficient staffing, high turnover, inadequate training, and poor communication—further reduce SOC effectiveness by straining human resources, eroding institutional knowledge, and inhibiting timely information sharing (Agyepong et al., 2020; Kokulu et al., 2019; Janos & Dai, 2018; Reeves & Ashenden, 2023; Onwubiko & Ouazzane, 2019; Onwubiko, 2017).

Addressing these vulnerabilities demands a holistic approach. Optimizing staffing levels, enhancing continuous training, and establishing clear communication protocols can strengthen SOC teams and restore operational continuity (Onwubiko, 2015), (Kokulu et al., 2019; Reeves & Ashenden, 2023; Onwubiko & Ouazzane, 2019). Technological improvements, such as integrating interoperable security platforms and adopting automation, can streamline workflows, alleviate analyst workloads, and improve incident detection and response times (Demertzis et al., 2018; Banati et al., 2022; E Degrass - US Patent App. 17/430 & 2022, 2022; Yuan & Zou, 2011). Moreover, advanced technologies like AI and ML can bolster human-machine

collaboration by rapidly analyzing massive data streams, identifying subtle threat patterns, and providing intelligent, data-driven insights without replacing the invaluable judgment and expertise of human analysts (Aslan & Samet, 2017; Demertzis et al., 2018).

While technical advancements are essential, they are not a substitute for a skilled and adaptable workforce. Cultivating a culture of continuous learning, innovation, and collaboration remains paramount. Although this review may not capture every industry insight and is subject to the evolving threat landscape, it reinforces that future research should focus on implementing AI and ML in real-world SOC settings and examining the interplay between technological enhancements and human factors. Ultimately, improving SOC resilience hinges on integrating technical refinements with strategic workforce development to safeguard organizations against increasingly sophisticated cyber threats.

Conclusion

This review has highlighted the interconnected technical and organizational vulnerabilities affecting SOCs and their substantial impact on cybersecurity. Issues such as tool misconfigurations, outdated systems, staffing shortages, high turnover rates, inadequate training, and poor communication channels collectively reduce the SOC's ability to detect and respond promptly to sophisticated threats. These findings underscore the need for a comprehensive approach that integrates both technology and human capital.

Strengthening SOC resilience involves investing in advanced technologies like AI and ML, which can process large datasets, identify threat patterns, and automate routine tasks. Simultaneously, organizations must address staffing challenges through better resource allocation, continuous training, and fostering collaborative work environments. Ensuring that analysts remain informed, supported, and engaged is vital for maintaining robust security operations.

As cyber threats evolve, future research should investigate how AI and ML can be practically integrated into SOCs, considering factors such as trust in automation, analyst learning curves, and potential shifts in job roles. Striking a balance between innovative technology, skilled personnel, and a supportive organizational culture can ensure that SOCs remain adaptive, proactive, and effective in safeguarding organizational assets and infrastructure.

REFERENCES

Agyepong, E., Cherdantseva, Y., Reinecke, P. & Burnap, P. (2020). *Cyber Security Operations Centre Concepts and Implementation*. https://doi.org/10.4018/978-1-7998-3149-5.ch006

Anton, Philip S., Anderson, Robert H., Mesic, Richard, Scheiern, & Michael. (n.d.). *The Vulnerability Assessment & Mitigation Methodology*. RAND NATIONAL DEFENSE RESEARCH INST SANTA MONICA CA. Retrieved November 26. — 2024, from https://apps.dtic.mil/sti/citations/ADA440505

Aslan, Ö., & Samet, R. (2017). Mitigating cyber security attacks by being aware of vulnerabilities and bugs. *Proceedings - 2017 International Conference on Cyberworlds, CW 2017 - in Cooperation with: Eurographics Association International Federation for Information Processing ACM SIGGRAPH*, *2017 — January*. https://doi.org/10.1109/CW.2017.22

Banati, A., Rigo, E., Fleiner, R., & Kail, E. (2022). Use cases of attack graph for SOC optimization purpose.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is.4. Number 24 (2025). Pp. 20–42
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.002

УДК 004.931

DEVELOPMENT OF A MODEL FOR SOIL SALINITY SEGMENTATION BASED ON REMOTE SENSING DATA AND CLIMATE PARAMETERS

G.B. Abdikerimova[1], M.B. Yessenova[*1], G. Yusupova[2], A.S. Tynykulova[3], A.M. Nedzved[4]

Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan;
[2]ALT UNIVERSITY, Almaty, Republic of Kazakhstan;
[3]Astana International University, Astana, Kazakhstan;
[4]Belarusian State University, Minsk, Republic of Belarus.
E-mail: moldir_11.92@mail.ru

Abdikerimova Gulzira — Associate Professor, Department of Information Systems, Eurasian National University named after L.N. Gumilyov, PhD, Astana, Kazakhstanhttps://orcid.org/0000-0002-4953-0737;
Yessenova Moldir — Senior Lecturer, Department of Information Systems, L.N. Gumilyov Eurasian National University, PhD, Astana, Kazakhstan
E-mail: moldir_11.92@mail.ru, https://orcid.org/0000-0002-2644-0966;
Yusupova Gulbahar — ALT UNIVERSITY. Departments «Radio Engineering and Telecommunications», Almaty, Kazakhstan
https://orcid.org/0000-0001-9765-2221;
Tynykulova Assemgul — Astana International University, PhD, Senior Lecturer at the Higher School of Information Technology and Engineering, Astana, Kazakhstan
https://orcid.org/o0000-0002-4557-6869;
Nedzved Alexander Mikhailovich — Doctor of Technical Sciences, Associate Professor, Professor of the Department of Computer Technologies and Systems, Faculty of Applied Mathematics and Informatics, Minsk, Belarus
https://orcid.org/0000-0001-6367-5900.

Abstract. The paper presents a hybrid machine learning model for the spatial segmentation of soils by salinity using multispectral satellite data from Sentinel-2 and climate parame-ters of the ERA5-Land model. The proposed method aims to solve the problem of ac-curate soil cover segmentation under climate change and high spatial heterogeneity of data. The approach includes the sequential application of unsupervised learning algo-rithms (KMeans, hierarchical clustering, DBSCAN),

the XGBoost model, and a multi-tasking neural network that performs simultaneous classification and regression. At the first stage, pseudo-labels are formed using KMeans, then a probabilistic assessment of object membership in classes and ensemble voting of clustering algorithms are car-ried out. The final model is trained on an extended feature space and demonstrates improved results compared to traditional approaches. Experiments on a sample of 33,624 observations (23,536 — training sample, 10,088 — test sample) showed an in-crease in the Silhouette Score value from 0.7840 to 0.8156 and a decrease in the Da-vies-Bouldin Score from 0.3567 to 0.3022. The classification accuracy was 99.99 %, with only one error in more than 10,000 test objects. The results confirmed the proposed method's high efficiency and applicability for remote monitoring, environmental analysis, and sustainable land management.

Keywords: soil salinity segmentation; remote sensing; ensemble clustering; Sentinel-2; ERA5-Land; XGBoost; multi-task neural network

For citation: G.B. Abdikerimova, M.B. Yessenova, G. Yusupova, A.S. Tynykulova, A.M. Nedzved. Development of a model for soil salinity segmentation based on remote sensing data and climate parameters. 2025. Vol. 6. No. 21. Pp. 20–42 (In Kaz.). https://doi.org/10.54309/IJICT.2025.21.1.002.

Conflict of interest: The authors declare that there is no conflict of interest.

ЖЕРДІҢ ТҰЗДЫҒЫН СЕГМЕНТАЦИЯЛАУ МОДЕЛІН ДАЛАЛЫҚ ЗЕРТТЕУЛЕР МЕН КЛИМАТТЫҚ ПАРАМЕТРЛЕР НЕГІЗІНДЕ ӘЗІРЛЕУ

Г.Б. Абдикеримова[1], М.Б. Есенова[1*], Г.М. Юсупов[2], А.С.Тыныкулова[3], А.М. Недзьведь[4]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2]ALT UNIVERSITY, Алматы, Қазақстан Республикасы;
[3]Астана Халықаралық университет, Астана, Қазақстан;
[4]Беларусь мемлекеттік университеті, Минск, Беларусь Республикасы.
E-mail: moldir_11.92@mail.ru

Абдикеримова Гульзира Бахытбековна – Л.Н. Гумилев атындағы Еуразия ұлттық университетінің Ақпараттық жүйелер кафедрасының доценті, PhD, Астана, Қазақстан, Сатпаев к., 2, 010000
https://orcid.org/0000-0002-4953-0737;
Есенова Молдир Балкаировна – Л.Н. Гумилев атындағы Еуразия ұлттық университетінің Ақпараттық жүйелер кафедрасының аға оқытушысы, PhD, Астана, Қазақстан
E-mail: moldir_11.92@mail.ru, https://orcid.org/0000-0002-2644-0966;
Юсупова Гульбахар Мадреймовна – LT UNIVERSITY. «Радиотехника және телекоммуникациялар» кафедрасының қауымдасқан профессоры, PhD,

Алматы, Қазақстан
https://orcid.org/0000-0001-9765-2221;
Тыныкулова Әсемгүл Сергожақызы — Астана Халықаралық университеті, PhD, аға оқытушы Ақпараттық технология және инженерия жоғарғы мектебі, Астана қаласы, Қазақстан
https://orcid.org/0000-0002-4557-6869;
Недьзвядь Александр Михайлович — т.ғ.д., доцент, Қолданбалы математика және информатика факультетінің Компьютерлік технологиялар және жүйелер кафедрасының профессоры, Минск қаласы, Беларусь
https://orcid.org/0000-0001-6367-5900.

Аннотация. Мақалада Sentinel-2 көпспектралды серіктік деректері мен ERA5-Land моделінің климаттық параметрлерін қолдана отырып, топырақтың минералды тұздығына негізделген кеңістік сегментациясына арналған машиналық оқытудың гибридтік моделі ұсынылған. Ұсынылған әдіс климаттың өзгеруіне және деректердің жоғары кеңістіктік гетерогенділігіне байланысты топырақ жамылғысын дәл сегментациялау мәселесін шешуге бағытталған. Бұл әдіс бақылаусыз оқыту алгоритмдерін (KMeans, иерархиялық кластерлеу, DBSCAN), XGBoost моделін және бір уақытта классификация мен регрессияны жүзеге асыратын көп тапсырмалы нейрон желісін қолдануды қамтиды. Алғашқы кезеңде KMeans арқылы жалған атаулар құрылады, содан кейін объектілердің сыныптарға жату ықтималдығын бағалау жүргізіледі және кластерлеу алгоритмдерінің біріктірілген дауысы қолданылады. Соңғы модель кеңейтілген белгі кеңістігінде оқытылып, дәстүрлі әдістермен салыстырғанда жақсартылған нәтижелер көрсетеді. 33 624 бақылаудан тұратын таңдамалар бойынша (23 536 — оқу таңдамасы, 10 088 — тест таңдамасы) Silhouette Score мәнінің 0,7840-тан 0,8156-ға дейін өсіп, Davies-Bouldin Score мәнінің 0,3567-ден 0,3022-ге дейін төмендегені көрсетілді. Сыныптау дәлдігі 99,99 %-ды құрады, 10 000 тест объектісінде тек бір қате жасалды. Нәтижелер ұсынылған әдістің қашықтан мониторинг жүргізу, қоршаған ортаны талдау және жер ресурстарын тұрақты басқару үшін жоғары тиімділігі мен қолданылу мүмкіндігін растады.

Түйін сөздер: топырақ тұздылығын сегментациялау; алыстан зондтау; ансамбль кластерлеу; Sentinel-2; ERA5-Land; XGBoost; көп міндетті нейрондық желі

Дәйексөздер үшін: Г.Б. Абдикеримова, М.Б. Есенова, Г.М. Юсупова, А.С.Тыныкулова, А.М. Недзьведь. Жердің тұздығын сегментациялау моделін далалық зерттеулер мен климаттық параметрлер негізінде әзірлеу// Халықаралық ақпараттық және коммуникалық технологиялар журналы. 2025. Т. 6. No. 21. 20–42 бет. (қазақ тілінде). https://doi.org/10.54309/IJICT.2025.21.1.002.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА МОДЕЛИ СЕГМЕНТАЦИИ ЗАСОЛЕННОСТИ ПОЧВ НА ОСНОВЕ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ И КЛИМАТИЧЕСКИХ ПАРАМЕТРОВ

Г.Б. Абдикеримова[1], М.Б. Есенова[*1], Г.М. Юсупов[2], А.С.Тыныкулова[3], А.М. Недзьведь[4]

Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан;

[2]ALT university, Алматы, Республика Казахстан;

[3]Международный университет Астана, Астана, Казахстан;

[4]Белорусский государственный университет, Минск, Республика Беларусь.

E-mail: moldir_11.92@mail.ru

Абдикеримова Гульзира Бахытбековна — доцент кафедры информационных систем Евразийского национального университета имени Л.Н. Гумилева, PhD, Астана, Казахстан

E-mail: gulzira1981@mail.ru, https://orcid.org/0000-0002-4953-0737;

Есенова Молдир Балкаировна — старший преподаватель кафедры информационных систем Евразийского национального университета имени Л.Н. Гумилева, PhD, Астана, Казахстан

https://orcid.org/0000-0002-2644-0966;

Юсупова Гульбахар Мадреймовна — Ассоцированный профессор кафедры «Радиотехники и телекоммуникации», PhD, Алматы, Қазақстан

https://orcid.org/-0000-0001-9765-2221;

Тыныкулова Асемгуль Сергужаевна — Международный университет Астана, PhD, старший преподаватель Высшей школы информационных технологии и инженерии, Астана, Казахстан

https://orcid.org/0000-0002-4557-6869;

Недзьведь Александр Михайлович — д.т.н, доцент, профессор кафедры Компьютерных технологии и систем, факультета Прикладной математики и информатики, Республика Беларусь, Минск

https://orcid.org/0000-0001-6367-5900.

Аннотация. В статье представлен гибридная модель машинного обучения для пространственной сегментации почв по солености с использованием многоспектральных спутниковых данных Sentinel-2 и климатических параметров модели ERA5-Land. Предложенный метод направлен на решение проблемы точной сегментации почвенного покрова в условиях изменения климата и высокой пространственной гетерогенности данных. Подход включает последовательное применение алгоритмов неконтролируемого обучения (KMeans, иерархическая кластеризация, DBSCAN), модели XGBoost и многозадачной

нейронной сети, которая выполняет одновременную классификацию и регрессию. На первом этапе формируются псевдоназвания с использованием KMeans, затем осуществляется вероятностная оценка принадлежности объектов к классам и объединяющее голосование алгоритмов кластеризации. Финальная модель обучается на расширенном пространстве признаков и демонстрирует улучшенные результаты по сравнению с традиционными подходами. Эксперименты на выборке из 33 624 наблюдений (23 536 — обучающая выборка, 10 088 — тестовая выборка) показали увеличение значения Silhouette Score с 0,7840 до 0,8156 и снижение значения Davies-Bouldin Score с 0,3567 до 0,3022. Точность классификации составила 99,99 %, было допущено только одна ошибка более чем в 10 000 тестовых объектов. Результаты подтвердили высокую эффективность и применимость предложенного метода для дистанционного мониторинга, анализа окружающей среды и устойчивого управления земельными ресурсами.

Ключевые слова: сегментация солености почвы; дистанционное зондирование; ансамблевое кластеризирование; Sentinel-2; ERA5-Land; XGBoost; многоцелевые нейронные сети

Для цитирования: Г.Б. Абдикеримова, М.Б. Есенова, Г.М. Юсупова, А.С.Тыныкулова, А.М. Недзьведь. Разработка модели сегментации засоленности почв на основе данных дистанционного зондирования и климатических параметров//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 21. Стр. 20–42. (На каз.). https://doi.org/10.54309/IJICT.2025.21.1.002.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе

Құқықтық процестерді автоматтандыру және цифрландыру (Karpenko et. al., 2023; Parycek et. al., 2023) көптеген елдерде, соның ішінде Қазақстанда да, сот жүйесін модернизациялаудың маңызды бөлігіне айналуда. Машиналық оқыту сияқты заманауи технологиялар құқық қолдану тәжірибесінің тиімділігін айтарлықтай арттыруға мүмкіндік береді, әсіресе нақты мәліметтерге шектеулі қолжетімділік жағдайында (Ruslan et. al., 2023; Said et. al., 2023). Сот шешімдерінің нәтижелерін деректер негізінде болжау сот процестерінде бар заңдылықтарды жақсырақ түсінуге, сондай-ақ уақыт пен ресурстарды барынша үнемдей отырып, негізделген шешімдер қабылдауға көмектеседі. Алайда, Қазақстанда сот істеріне қатысты нақты деректерге қолжетімділік құпиялылыққа байланысты шектеулі, бұл ғылыми талдау жүргізуге және олардың негізінде болжам жасау модельдерін әзірлеуге айтарлықтай кедергілер туғызады (Bianchini et. al., 2024; Setiawan et. al., 2024). Осы мәселені шешу Топырақтың тұзденуі – ауыл шаруашылығы дақылдарына, экожүйелердің жұмысына және аймақтардың су режиміне тікелей әсер ететін негізгі экологиялық проблемалардың бірі (Aksoy

et all., 2024; Bo et all 2024). Климаттың өзгеруі және ауылшаруашылық жерлерін пайдаланудың интенсификациясы жағдайында топырақтың деградация дәрежесі (Bougiouklis et all., 2025) керемет жылдам қарқынмен артады. Сондықтан тұздылықтың тиімді мониторингі мен болжау құралдарын әзірлеу қажет. Спутниктік деректер, атап айтқанда, Sentinel-2 платформаларынан алынған мульти-диапазонды суреттер (Cui et all., 2025), ERA5-Land сияқты климаттық модельдермен біріктіріліп, топырақ жамылғысының күйі туралы кең ақпарат береді. Дегенмен, оларды өңдеу деректерді талдаудың заманауи әдістерін қажет етеді. Қазіргі уақытта ғылыми әдебиеттерде NDVI, NDWI және NDSI (Gao et all., 2025; Gao et all., 2025), сияқты спектрлік индекстерді қоса алғанда, топырақтың тұздылығын бағалаудың әртүрлі тәсілдері, сонымен қатарты спектрлік и (Geoderma et all., 2024) және регрессия (Jia et all., 2024; Jiang et all.,2025) әдістерін қолданатын машиналық оқыту ұсынылған. Дегенмен, бақыланатын оқытуға негізделген классикалық әдістер сенімді деректер белгілерінің болмауына және мүмкіндіктерді таратудың гет-эрогенділігіне байланысты бірнеше шектеулерге тап болады. Осыған байланысты, соңғы жылдары болжау дәлдігін жақсарту үшін бақыланбайтын алгоритмдерді, ансамбльдік әдістерді және терең нейрондық желілерді біріктіретін гибридті әдістерге қызығушылық артты. Бұл саладағы өзекті мәселелердің бірі әр түрлі топырақ түрлерінің, соның ішінде сортаң, қалыпты және құмды топырақтардың спектрлік ерекшеліктерін ескеру қажеттілігі болып табылады, өйткені құмды топырақтардың спектрлік сипаттамалары сортаңданған аймақтардан келетін сигналдармен қабаттасуы мүмкін. Сонымен қатар, маңызды міндет - жоғары кеңістіктік және спектрлік өзгергіштік жағдайында деректерді бейімдеуге болатын алгоритмдерді әзірлеу. Бұл жұмыста біз бақыланбайтын алгоритмдерді (KMeans (Kaliyeva et all., 2025; Land et all., 2024), агломеративті кластерлеу, DBSCAN (Liu et all., 2024; Luo et all., 2025)) басқарылатын әдістермен (XGBoost (Lusiana et all., 2025., Lusiana et all.,2025), көп тапсырмалы нейрондық желі) біріктіретін гибридті модельді ұсынамыз. Бастапқы кезең жалған белгілерді қалыптастыру үшін KMeans әдісін пайдаланып деректерді кластерлеуді қамтиды, содан кейін олар нақты сыныптарға жататын нысандардың ықтималдығын есептейтін XGBoost үлгісін үйрету үшін пайдаланылады. Әрі қарай, бірнеше кластерлік алгоритмдердің дауыс беруімен ансамбльдік тәсіл қолданылады, содан кейін түпкілікті белгілер топырақтың тұздылығын жіктеу және регрессиялау мәселелерін бір уақытта шешетін көп тапсырмалы нейрондық желіні оқыту үшін пайдаланылады.

Зерттеу ұсынылып отырған әдіс дәстүрлі кластерлеу және жіктеу әдістерімен салыстырғанда топырақты тұздылық бойынша дәлірек сегменттеуді қамтамасыз етеді деп болжайды. Эксперименттерде модель сапасының негізгі көрсеткіштері бағаланады, соның ішінде Sil-houette Score, Davies-Bouldin Score, Calinski-Harabasz Score және гибридті ансамбльдің соңғы болжамды үлгінің сапасына әсері. Зерттеудің негізгі нәтижелері гибридті тәсілді қолдану тұзды әсер ететін топырақ сегментациясының дәлдігін жақсартады, әртүрлі топырақ

жамылғысының түрлері арасындағы шекаралық әсерлердің әсерін азайтады және модельді деректердің кеңістіктегі біркелкі еместігіне бейімдейді. Бұл зерттеу топырақты қашықтықтан зондтау әдістерін жақсартуға көмектеседі және тұзданған аймақтағы жердің деградациясын бақылау үшін пайдаланылуы мүмкін. Топырақтың тұзденуы – егістік алқаптарының өнімділігіне, экожүйе функцияларына және суға үлкен әсер ететін экологиялық маңызды мәселе. Климаттың өзгеруі және ауыл шаруашылығының қарқынды дамуы туралы, топырақтың тұзденуін бақылау және болжау құралдарын құрудың өзектілігі біртіндеп артып келеді. Ғалымдар соңғы жылдары топырақтың тұздылығын болжау мен картаның дәлдігін жақсарту үшін спутниктік бейнелеуді, машиналық оқыту алгоритмдерін және географиялық ақпараттық жүйелерді біріктіруге тырысты. Зерттеудің жаңа бағыттарының бірі топырақтың тұздылығын анықтау үшін қашықтықтан зондтау және спектрлік көрсеткіштерді пайдалану болып табылады. Зерттеуде (Paramonova et all., 2025) Sentine l-2 деректері мен терең оқыту алгоритмдері арқылы тұзды жерлерді автоматты түрде анықтау әдістері талданды. Нәтижелер спектрлік индекстерді (NDVI, NDSI, BI) көпқабатты нейрондық желілермен біріктіру тұзды топырақтарды жоғары классификациялық дәлдікке (91,2 % дейін) мүмкіндік беретінін көрсетті. Басқа зерттеу (Sadyrova et all., 2025) топырақ жағдайын талдау үшін гиперспектрлік деректерді пайдаланудың жетілдірілген әдістемесін ұсынды. Тұздылық өзгерістерінің кеңістік-уақыттық заңдылықтарын дәлірек модельдеуге мүмкіндік беретін деректер трансформаторлық әдістерді қолдану арқылы өңделді. Сонымен қатар, топырақтың тұздану динамикасын болжау үшін климаттық деректер мен спутниктік суреттерді біріктірудің әртүрлі тәсілдері зерттелуде. Мысалы, жұмыс (Tashpolat et all., 2025) топырақ температурасы мен ылғалдылығының өзгеруін ескере отырып, топырақтың деградациясының өзгеруін болжауға мүмкіндік беретін машиналық оқытуға және ERA5-Land моделінің деректеріне негізделген әдісті ұсынады. Градиентті күшейту және кездейсоқ орман сияқты жиынтық әдістер дәстүрлі сызықтық регрессия үлгілерімен салыстырғанда болжау дәлдігін 15%-ға арттырды. Зерттеудің тағы бір маңызды бағыты топырақтың сортаңдануына байланысты ауылшаруашылық жерлерінің деградациясын бағалау болып табылады. Зерттеу (Tussupov et all., 2024) аридті аймақтардағы топырақтың тұздану дәрежесін сандық түрде анықтауға мүмкіндік беретін мультиспектрлік зондтау және машиналық оқытуға негізделген тәсілді ұсынады. Авторлар бақыланбайтын әдістерді (KMeans кластерлеу, DBSCAN) бақыланатын әдістермен (XGBoost, нейрондық желілер) біріктіру классификация қателерін азайтуға және топырақ сипаттамаларындағы күрделі сызықтық емес тәуелділіктерді қарастыруға мүмкіндік беретінін атап көрсетеді.

Қашықтықтан зондтаумен қатар, жергілікті өлшемдерге және тұздану процестерін модельдеуге негізделген тәсілдер белсенді түрде әзірленуде. Зерттеуде (Tussupov et all., 2024) Топырақтың тұздылығын картаға түсіру үшін экипажсыз ұшу аппаратының (ҰҰА) деректерін геостатистикалық модельдермен

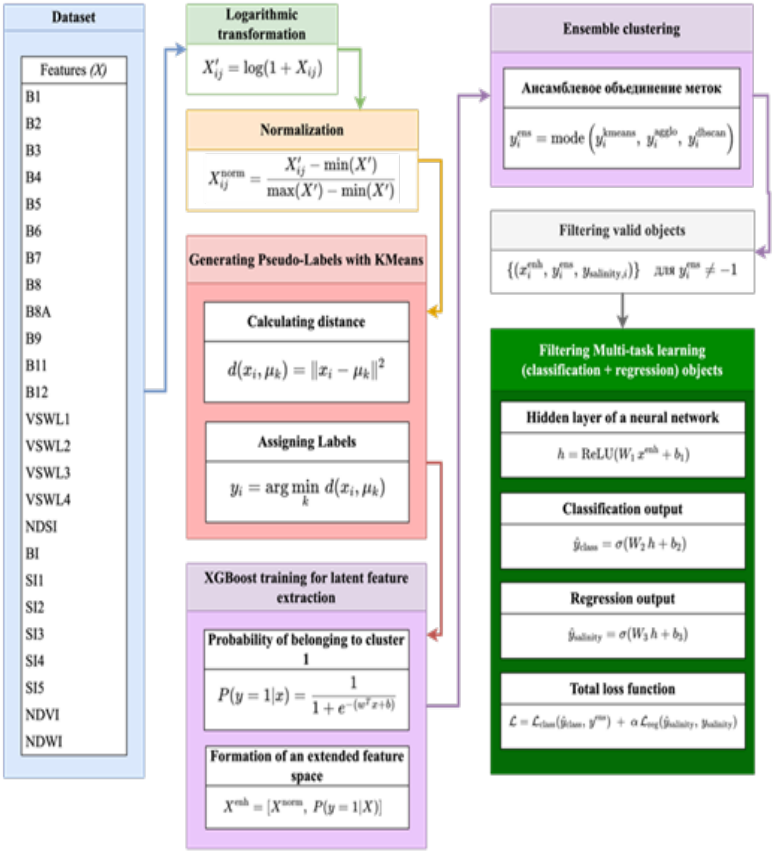
біріктіру мүмкіндіктері талданды. Жұмыс ұшқышсыз ұшу аппараттары терең оқыту әдістерімен үйлесе отырып, жоғары кеңістіктік дәлдікті қамтамасыз ете отырып, бақылау шығындарын азайта алатынын көрсетеді. Соңында, зерттеу (Velilla et all., 2025) әртүрлі аймақтардағы топырақ тұздылығын болжау үшін спутниктік бақылау, жердегі өлшеулер және машиналық оқытуды біріктіретін кешенді тәсілді қарастырады. Авторлар ансамбльдік үлгілерді пайдалану, әсіресе далалық деректердің шектеулі қолжетімділігі жағдайында болжамдардың сенімділігін арттыратынын атап көрсетеді. Осылайша, қазіргі заманғы зерттеулер топырақтың тұзденуін бақылау және болжауда айтарлықтай прогресті көрсетеді. Қашықтықтан зондтау деректерін, климаттық модельдерді және машиналық оқыту алгоритмдерін біріктіру тұзданған жерлерді жіктеудің дәлдігін арттыруға және топырақ қасиеттерінің кеңістік-уақыттық динамикасына негізделген болжамды модельдерді жасауға мүмкіндік береді. Бұл жерді тұрақты басқару, топырақтың деградациясының әсерін азайту және ауыл шаруашылығы өндірісінің тиімділігін арттыру үшін қажет.

Әдістер мен материалдар

Зерттеуде көпжолақты Sentinel-2 спутниктік суреттері және ERA5-Land климаттық моделі қолданылды. Sentinel-2 спектрлік деректеріне B2, B3, B4, B8, B8A, B11, B12 арналары, сондай-ақ алынған спектрлік индекстер (NDVI, NDWI, NDSI, SI1–SI5, BI) кірді, бұл өсімдік жамылғысын, топырақ ылғалдылығын және жер бетінің минералдану дәрежесін талдауға мүмкіндік берді. Топырақ суының көлемдік көрсеткіштері (VSWL1–VSWL4) ERA5-Land моделінен төрт қабатқа (0–7 см, 7–28 см, 28–100 см, 100–289 см) алынды, бұл топырақтың сортаңдануына ылғалдың әсерін есепке алуға мүмкіндік береді. Талдау үшін әртүрлі топырақ сипаттамалары бар үш аймақ таңдалды: тұзданудың жоғары деңгейімен сипатталатын Арал теңізінің бұрынғы түбі; Тұздылығы қалыпты және өсімдіктері типтік Бозайғыр ауылы (Ақмола облысы, Қазақстан); және құмды топырақтардың спектрлік сипаттамаларын бағалау үшін бақылау аймағы ретінде пайдаланылған ең аз өсімдік жамылғысы бар құмды жерлер. Деректер 2019–2024 жылдар аралығын қамтыды, топырақтың тұздану заңдылықтарын және олардың уақыт бойынша динамикасын анықтау үшін жеткілікті ақпарат береді.

1-суретте машиналық оқыту әдістерін біріктіретін топырақ тұздылығын кластерлеуге және болжауға осы тәсілдің әдістемесі суреттелген. Ағындық диаграмма алдын ала өңдеуден және жалған белгіні жасаудан бастап ансамбльді оқытуға және көп тапсырмалы нейрондық модельдеуге дейінгі деректерді өңдеудің дәйекті кезеңдерін көрсетеді.

1. Зерттеудің бірінші кезеңі деректерді алдын ала өңдеуден тұрады. Бұл кезеңде бастапқы деректер одан әрі талдау үшін дайындалады, соның ішінде үлестірімнің асимметриясын жою және мүмкіндіктерді бір шкалаға келтіру. Біріншіден, логарифмдік түрлендіру қолданылады, ол шеткі мәндердің әсерін және үлестірімнің қисаюын азайтады.



Сур. 1. *Мәліметтерді өңдеу кезеңдері және топырақтың тұздылығын болжау моделін құру.*

Әрі қарай деректер қалыпқа келтіріледі, сондықтан барлық мүмкіндіктер [0,1] ауқымында мәндерге ие болады. Бұл келесі алгоритмдердің конвергенциясын жақсартады және әрбір мүмкіндіктің модельге біркелкі әсер етуін қамтамасыз етеді.

Логарифмдік түрлендіру (2):

$$X_{ij}^{'} = \log\left(1 + X_{ij}\right) \tag{1}$$

where X_{ij}– is the original value of the j - th feature of objecti, where $X \in Rn \times d$, X'_{ij} – prime is the transformed value after applying the logarithm, $X' \in R_n \times d$.

2. Нормализация (2):

$$X_{ij}^{\text{norm}} = \frac{X'_{ij} - \min(X')}{\max(X') - \min(X')} \tag{2}$$

мұндағы X'_{ij}– жай – логарифмдік түрлендіруден кейінгі мүмкіндік матрицасы,

X_{ij}^{norm} norm– нормаланған мүмкіндік матрицасы, мұнда мәндер 0-ден 1-ге дейінгі диапазонға жеткізіледі.

Деректерді алдын ала өңдеуден кейін псевдобелгілер KMeans алгоритмі арқылы жасалады. Бұл кезеңде KMeans алгоритмі X_{norm} нормаланған деректерді екі кластерге бастапқыда бөлу үшін қолданылады. Алгоритм кластер орталықтарын есептеп, ең жақын орталыққа дейінгі ең аз қашықтық негізінде әрбір нүктеге белгіні тағайындайды. Алынған жалған белгілер бастапқы деректерді бөлуді (3) және белгіні тағайындауды (4) қамтамасыз ететін XGBoost моделін одан әрі оқыту үшін нұсқаулық ретінде пайдаланылады.

$$d(x_i, \mu_k) = |x_i - \mu_k \tag{3}$$

мұндағы x_i– $X_{norm}j$ - объектінің ерекшелік векторы, μ_k– кластерінің центрі, мұндағы k =1,2, $d(x_i, \mu_k)$– x_i и μ_k арасындағы евклидтік қашықтықтың квадраты.

$$y_i = \arg\min_k d(x_i, \mu_k) \tag{4}$$

мұндағы $d(x_i, \mu_k)$– әрбір x_i объектісінің кластер орталықтарына дейінгі есептелген қашықтығы, k – кластер индексі, $k \in \{1,2\}$, y_i – i - нысанға тағайындалған жалған белгі, $Y \in \{0,1\}$.

Бұл қадамда XGBoost белгілі бір кластерге мүшеліктің ықтималдық бағалауларын алу үшін Y жалған белгілермен X^{norm} нормаланған деректері бойынша оқытылады. Бұл ықтималдықтар қарапайым кластерлеу әдістерімен түсірілмеген деректердегі жасырын үлгілерді көрсетеді. Нәтиже (ықтималдық $(y = 1 \mid x)$ (5) X^{enh} (6) кеңейтілген мүмкіндік кеңістігін құра отырып, бастапқы мүмкіндіктерге қосылады).

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \tag{5}$$

Мұндағы x – X^{norm} ерекшелік векторы, w – модель салмақтарының векторы, b – (қиғаштық), $P(y = 1 \mid x)$класқа жататын x объектінің ықтималдығы, мән [0,1] диапазонында.

$$X^{enh} = X^{norm}, (y = 1 \mid X) \tag{6}$$

Мұндағы X^{norm} – нормаланған деректер, $P(y = 1 \mid X)$– формула (6) бойынша алынған әрбір нысан үшін ықтималдық векторы, X^{enh} – бастапқы мүмкіндіктерді және қосымша ықтималдық мүмкіндігін қамтитын кеңейтілген мүмкіндік матрицасы, $X^{enh} \in R_n \times (d+1)$.

Келесі қадам - ансамбльдік кластерлеуді қолдану. Бұл қадамда X^{enh} кеңейтілген мүмкіндік кеңістігі үш кластерлеу әдісін қатар қолдану үшін пайдаланылады: KMeans, агломеративті кластерлеу және DBSCAN. Әрбір әдіс нысандарға белгілерді дербес тағайындайды, содан кейін нәтижелер дауыс беру процедурасы арқылы біріктіріледі, мұнда нысан үшін соңғы белгі ең жиі кез-

десетін мән ретінде анықталады. Бұл ансамбльдік тәсіл әрбір әдістің жеке қателерінің әсерін азайтады және деректерді неғұрлым тұрақты және дәлірек бөлуді қамтамасыз етеді (7).

$$y_i^{\text{ens}} = \text{mode}\left(y_i^{\text{kmeans}}, y_i^{\text{agglo}}, y_i^{\text{dbscan}}\right) \tag{7}$$

Мұндағы y_i^{kmeans} – нысанға тағайындалған белгі i алгоритмі Kmeans, y_i^{agglo} – нысанға тағайындалған белгі i агломеративті кластерлеу, y_i^{dbscan} – нысанға тағайындалған белгі i DBSCAN алгоритмі, y_i^{ens} – $Y_{ens} \in R_n$ нысанының соңғы ансамблі.

Ансамбльді кластерлеуден кейін кейбір нысандар «шу» (−1) белгісін алуы мүмкін, сондықтан әрі қарай оқыту үшін жарамды белгілері бар нысандар ғана таңдалады. Көп тапсырмалы үлгінің кірісі x^{enh} кеңейтілген мүмкіндіктері ((таңдалған нысандар үшін), сондай-ақ екі мақсатты айнымалы болады: (1) жіктеу үшін $y\text{^}ens$ класы және (2) тұздылықтың жалған белгісі $y^{salinity}$ біздің жағдайда 1-сынып үшін XGBoost қайтарған ықтималдық немесе басқа қолжетімді тұздылық (8).

$$\left\{\left(x_i^{\text{enh}}, y_i^{\text{ens}}, y_{\text{salinity},i}\right)\right\} \quad \text{для } y_i^{\text{ens}} \neq -1 \tag{8}$$

ұнда x^{enh} – кеңейтілген мүмкіндіктер, y^{ens} – ансамбль белгілері, $y^{salinity}$ – тұздылықтың псевдобелгілері (немесе дәл белгілер).

Қорытынды кезең – жіктеу мен регрессияны біріктіретін көп тапсырмалы оқыту. Бір уақытта екі мәселені шешетін нейрондық желі пайдаланылады: ансамбль белгілері бойынша жіктеу y^{ens} $y_{salinity}$ псевдобелгілері бойынша регрессия. Модельде стандартты кіріс қабаты, бірнеше жасырын қабаттар (9), содан кейін екі «бас» - классификацияға арналған шығыс (10) және регрессияға арналған шығыс (11). Жаттығу кезінде екі мәселе бойынша жоғалту функциялары (12) бірге ескеріледі.

$$h = \text{ReLU}\left(W_1 x^{\text{enh}} + b_1\right) \tag{9}$$

мұндағы x^{enh} – бір нысан үшін кеңейтілген мүмкіндік векторы, W_1, b_1– бірінші жасырын қабаттың салмақтық матрицасы және қиғаштық векторы, h – жасырын қабаттың белсендірілуі (жаңа мүмкіндікті көрсету).

$$\hat{y}_{class} = \sigma(W_2 h + b_2) \tag{10}$$

мұндағы h – жасырын қабат шығысы, W_2, b_2– классификациялық деңгейдің параметрлері, $\sigma(z) = \frac{1}{1+e^{-z}}$, $\hat{y}_{class} \in [0,1]$ – 1 классқа жататын болу ықтималдығы.

$$\hat{y}_{\text{salinity}} = \sigma(W_3 h + b_3) \tag{11}$$

мұндағы h – жасырын қабаттың шығысы, W_3, b_3 – регрессия қабатының параметрлері, $\sigma(z)$– мұнда болжамды [0,1] диапазонына әкелетін шектеуші ретінде пайдаланылады, $\hat{y}_{\text{salinity}} \in [0,1]$– болжамды тұздылық мәні.

$$\mathcal{L} = \mathcal{L}_{class}(\hat{y}_{class}, y^{\text{ens}}) + \alpha \mathcal{L}_{reg}(\hat{y}_{\text{salinity}}, y_{\text{salinity}}) \quad (12)$$

мұндағы $\hat{y}_{class}$, $\hat{y}_{class}$ – жіктеу үшін болжамды және нақты белгі$\hat{y}_{\text{salinity}} y_{salinity}$ – болжамды және нақты (немесе жалған) тұздылық белгісі, α – регрессия мәселесінің жалпы қатеге қосқан үлесін анықтайтын салмақ тұрақтысы, L – жаттығуда қолданылатын соңғы жоғалту функциясы.

Іске асыру үшін төмендегі бағдарламалық құрал пайдаланылды: нейрондық желілер үшін Python 3.9.12, TensorFlow 2.12.0, градиентті күшейту үшін XGBoost 1.6.2, кластерлеу және қалыпқа келтіру үшін Scikit-learn 1.1.3, геодеректерді өңдеу үшін GDAL 3.4.0, Matplot2 және визуализация үшін Sea.5.1.2. Спутниктік кескін деректері Google Earth Engine API арқылы алынды және өңделді. Модельдер жоғары есептеу жылдамдығы мен үлкен деректер сыйымдылығын қамтамасыз ету үшін NVIDIA RTX 3090 графикалық картасы (24 ГБ VRAM) және 128 ГБ жады бар есептеу серверінде орындалды. Sentinel-2 деректері пайдаланылған спутниктік деректерге Коперник ашық қолжетімділік хабы (https://scihub.copernicus.eu) арқылы қол жеткізіледі. ERA5-Жер климаты деректеріне ECMWF Climate Data Store (https://cds.climate.copernicus.eu) арқылы қол жеткізуге болады. Осылайша, ұсынылған әдісті басқа аймақтар үшін қайта шығаруға және бейімдеуге болады, бұл оны топырақтың тұздылығын талдаудың әмбебап құралына айналдырады.

Нәтижелер және талқылау.

Ұсынылған әдіс топырақтың тұзденуін болжау мен кеңістіктік сегменттеуде жоғары тиімділікті көрсетті. Тек KMeans алгоритміне негізделген негізгі тәсілден айырмашылығы, ансамбльді кластерлеуді, ықтималды мүмкіндіктерді байытуды және көп тапсырмалы нейрондық желінің архитектурасын біріктіретін әзірленген жүйе барлық негізгі көрсеткіштерде айтарлықтай жақсартуларды көрсетті. Бұл қашықтан бақылау тапсырмаларына тән күрделі гетерогенді деректерді өңдеу кезінде машиналық оқыту әдістерін біріктірудің орындылығын растайды. Модель екі қосымша деректер көзіне негізделген: Sentinel-2 мультиспектрлі кескіні және ERA5-Land моделінен алынған климаттық деректер. Біз тұзденуға әсер ететін беттік шағылысу қасиеттерін де, гидрологиясын да түсіру үшін екеуін де қолдандық. Коперник бастамасы бойынша Еуропалық ғарыш агенттігінің Sentinel-2 спутниктік жүйесі спектрдің үлкен бөлігінде (10–60 м) жоғары кеңістіктік ажыратымдылығы бар бейнелеуді ұсынатын MSI мультиспектрлі сенсорын алып жүреді. Талдау 10 метрлік ажыратымдылық диапазондарын пайдаланды: B2 (көк), B3 (жасыл), B4 (қызыл), B8 және B8A (қысқа толқынды инфрақызыл) және B11 және B12 (орта инфрақызыл). Бұл арналар өсімдік жамылғысы, ылғалдылық және жер бетінің шағылыстыру сипаттамалары туралы ақпаратты, соның ішінде минералды құрамы мен ықтимал тұздылық туралы кеңестерді береді. Модель екі қосымша

ақпарат көзіне негізделген: Sentinel-2 мультиспектрлі бейнелеу және климаттық ERA5-Жер үлгісі деректері. Біз бетінің шағылысу қасиеттерін және тұздануға әсер ететін гидрологияны ескеру үшін қолдандық. Еуропалық ғарыш агенттігінің «Коперник» бағдарламасының Sentinel-2 спутниктік жүйесінде спектрдің кең бөлігінде (10–60 м) жоғары ажыратымдылықты бейнелеуі бар мультиспектрлік сенсор MSI бар. Талдау 10 метрлік ажыратымдылық жолақтарын пайдаланды: B2 (көк), B3 (жасыл), B4 (қызыл), B8 және B8A (қысқа толқынды инфрақызыл) және B11 және B12 (ортаңғы инфрақызыл). Бұл деректер тұзды тік тасымалдау процестерін бағалау үшін қажет, яғни тереңдік пен маусымның функциясы ретінде шаймалау немесе жиналу. Негізгі спектрлік және гидрологиялық қасиеттерден басқа, модельдің өсімдіктер мен топырақтың ылғалдылық қасиеттеріне сезімталдығын арттыратын туынды спектрлік индекстер (SI1–SI5, NDVI, NDWI, NDSI, BI) болды. Осылайша, мүмкіндіктер жиынтығы тұздылықты модельдеуге кешенді тәсілді қамтамасыз ететін беткі және терең параметрлерді қамтиды. Бұл талдау мүмкіндіктерін кеңейтеді және әртүрлі табиғи-климаттық жағдайларда қолданғанда әдістің әмбебаптығын арттырады. Деректер Google Earth Engine арқылы алынды. Бұл жұмыста келесі көрсеткіштер пайдаланылды:
1.Нормалданған дифференциалды тұздылық индексі (NDSI) (13):

$$NDSI = \frac{R-NIR}{R+NIR} \tag{13}$$

мұндағы R – қызыл арнаның мәні (B4), NIR – жақын инфрақызыл арна (B8). Бұл көрсеткіш топырақтың тұздану деңгейін бағалау үшін қолданылады.
2. Жарықтық индексі (BI) (14):

$$BI = \sqrt{R^2 + NIR^2} \tag{14}$$

Бұл көрсеткіш беттің жарықтығын көрсетеді және топырақ құрылымын бағалауға көмектеседі.
3. Тұздылық индексі 1 (SI1) (15):

$$SI1 = \sqrt{B + R} \tag{15}$$

мұндағы B – көк арнаның мәні (B2).
4. Тұздылық индексі 2 (SI2) (16):

$$SI2 = \sqrt{G^2 + R^2 + NIR^2} \tag{16}$$

мұндағы G – жасыл арнаның мәні (B3).
5. Тұздылық индексі 3 (SI3) (17):

$$SI3 = \frac{SWIR1-(SWIR1 \cdot SWIR2)}{SWIR1} \tag{17}$$

мұндағы $SWIR1$ және $SWIR2$ – орташа инфрақызыл арналардың мәндері (B11 және B12).

6. Тұздылық индексі 4 (SI4) (18):

$$SI4 = \sqrt{R \cdot NIR} \tag{18}$$

7. Тұздылық индексі 5 (SI5) (19):

$$SI5 = \frac{R}{G \cdot \frac{R}{B}} \tag{19}$$

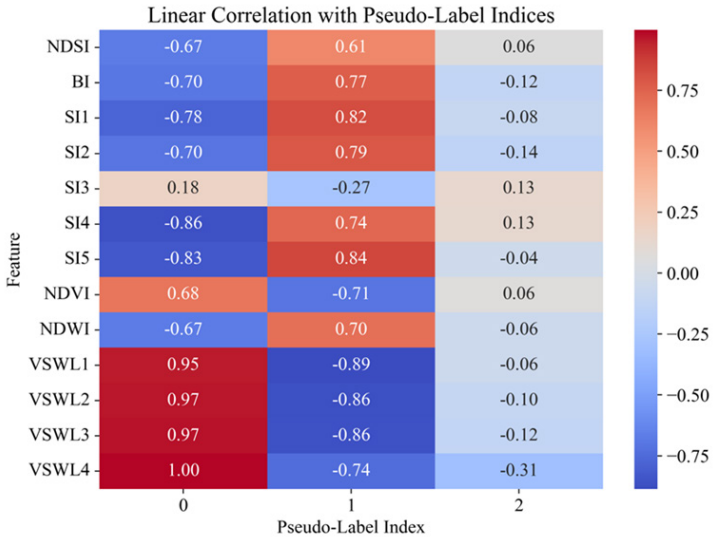
8. Өсімдік жамылғысының нормаланған айырмашылық индексі (NDVI) (20):

$$NDVI = \frac{NIR - R}{NIR + R} \tag{20}$$

9. Судың нормаланған айырма индексі (NDWI) (21):

$$NDWI = \frac{G - NIR}{G + NIR} \tag{21}$$

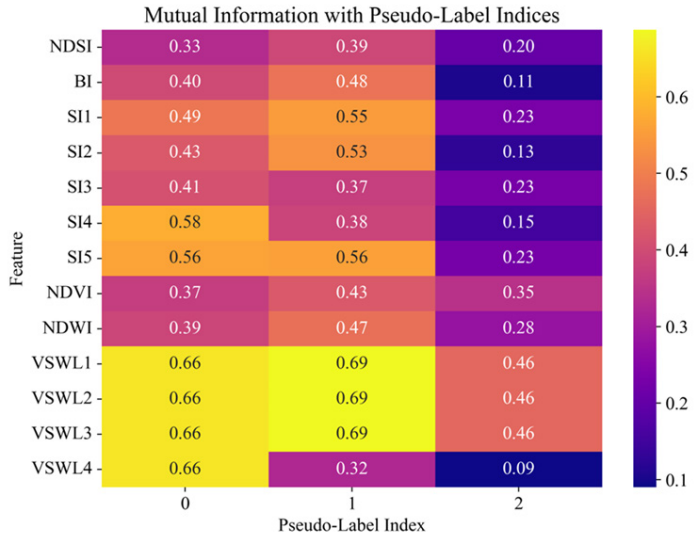
Олар топырақты бағалауда басты орынды алады, өйткені олар оның күйінің әртүрлі аспектілерін, соның ішінде тұздылық пен су балансын анықтайды. Зерттеу жалпы 33 624 деректер жолын өңдеді. Оның ішінде 23 536 жол модельді үйрету үшін пайдаланылды, ал қалған 10 088 жол кластерлеу сапасын сынау және бағалау үшін пайдаланылды. 2-суретте кластерлеуден алынған кластерлердің белгілері мен псевдобелгілері арасындағы сызықтық корреляцияның жылу картасы (Псевдо-белгі индексі) көрсетілген. Y осі спектрлік индекстерді (NDSI, NDVI, NDWI, SI1–SI5, BI) және гидрологиялық ERA5-Жер үлгісінің айнымалы мәндерін (VSWL1–VSWL4) қоса алғанда, пайдаланылған мүмкіндіктерді көрсетеді, ал X осі кластерлік индекстерді (0, 1, 2) көрсетеді. Түс шкаласы корреляция деңгейін білдіреді: Оң мәндер қызыл реңктермен көрсетіледі, бұл тікелей қатынасты көрсетеді, ал теріс мәндер көк түспен көрсетіледі, бұл кері қатынасты білдіреді. 0 кластерімен ең күшті оң корреляция барлық гидрологиялық айнымалылар үшін (VSWL1–VSWL4) байқалады, VSWL4 (1,00) бұл кластер түрімен топырақ ылғалдылығының жоғары деңгейінің байланысын көрсетеді. Сонымен қатар, 1 кластер үшін көптеген белгілер (атап айтқанда, SI1, SI2, SI5, NDVI) оң корреляцияны көрсетеді, бұл тұздылықтың спектрлік ерекшеліктерін көрсетуі мүмкін. 2-кластер әлсіз немесе шамалы теріс корреляциямен сипатталады, бұл оның аралық немесе аралас сипатын көрсетуі мүмкін. Мұндай визуализация кластер құруға жеке белгілердің үлесін жақсырақ түсінуге және кластерлеу кезінде анықталған әртүрлі топырақ түрлерінің физикалық қасиеттерін түсіндіруге мүмкіндік береді.



Сур. 2. Кластерлердің белгілері мен псевдобелгілері арасындағы сызықтық корреляцияның жылу картасы.

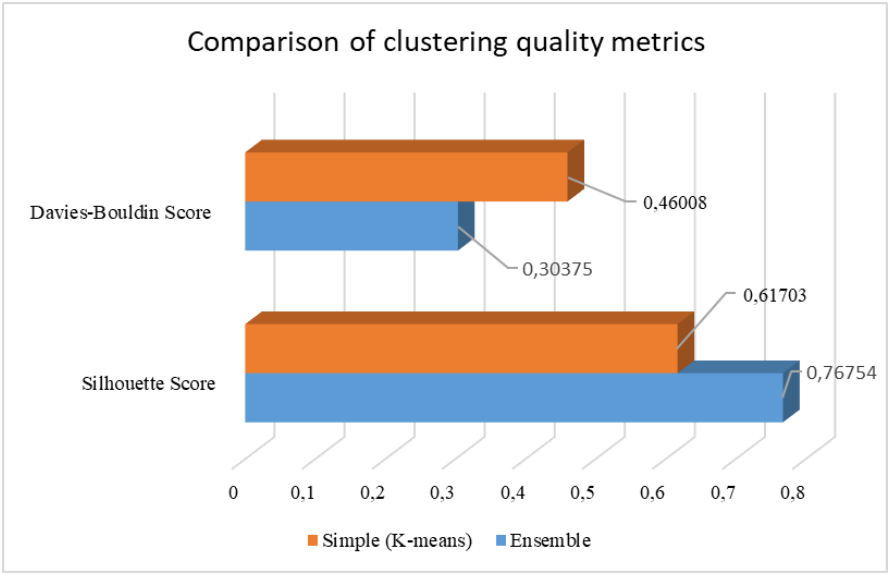
3-суретте кластерлеу процесінде алынған кластерлердің белгілері мен псевдобелгілерінің индекстері арасындағы өзара ақпараттың жылу картасы көрсетілген. Тік ось спектрлік индекстерді (NDSI, SI1–SI5, NDVI, NDWI, BI) және гидрологиялық параметрлерді (VSWL1–VSWL4) қамтитын мүмкіндіктерді, ал көлденең ось кластер индекстерін (0, 1, 2) көрсетеді. Түс шкаласы өзара ақпараттың мәнін көрсетеді, мұнда ашық сары реңктер кластер мүшелігіне қатысты мүмкіндіктің ақпараттылығының жоғары деңгейіне сәйкес келеді. 0 және 1 кластерлері үшін VSWL1, VSWL2 және VSWL3 гидрологиялық параметрлері (0,69) топырақ ылғалдылығы деңгейі бойынша деректерді бөлуге қатысты олардың маңызды мәнін растайтын максималды өзара ақпараттық мәндерге жетеді. Спектрлік сипаттамалардың ішінде SI4, SI5 және SI1 таңбалары бар ең жоғары өзара ақпаратқа ие, әсіресе 0 кластері мен 1 кластеріне қатысты, сондықтан топырақтың тұздылығы мен құрылымын анықтауда ең маңызды болып табылады. 2-кластермен салыстырғанда әрбір мүмкіндік үшін өзара ақпарат аз, бұл 2-кластердің құрылымының азырақ немесе аралас сипатын көрсетуі мүмкін. Осылайша, өзара ақпаратты визуализациялау бізге кластер құрылымын құруға ықпал ететін ең маңызды мүмкіндіктерді көруге мүмкіндік береді және қосымша мүмкіндіктерді таңдау және үлгіні түсіндіру үшін пайдаланылуы мүмкін.

4-суретте екі үлгінің кластерлік сапасы салыстырылады: біреуі KMeans алгоритмімен қарапайымырақ (қызғылт сары жолақтар) және бірнеше әдістерді біріктіретін ансамбльдік үлгі (көк жолақтар). Екі ішкі бағалау көрсеткіші пайдаланылды: Силуэт бағасы және Дэвис-Боулдин бағасы.



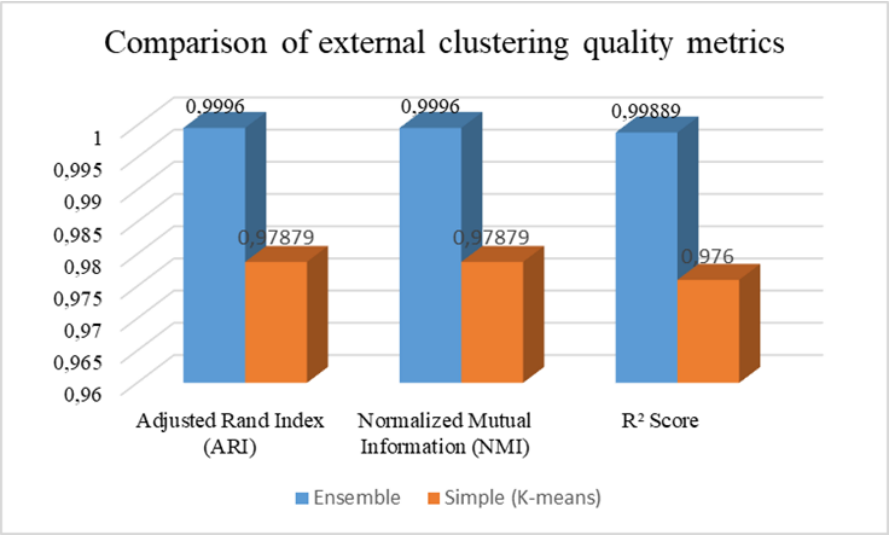
Сур. 3. Кластерлердің белгілері мен псевдобелгілері арасындағы өзара ақпараттың жылу картасы

Кластерлердің қаншалықты оқшауланғанын сипаттайтын «Силуэт» көрсеткіші қарапайым үлгіге (0,61703) қарағанда ансамбльдік үлгіде (0,76754) жоғары болды, бұл нысандар кластерлері арасындағы ұлғайтылған бөлуді көрсетеді. Сонымен қатар, Da-vies-Bouldin Score, неғұрлым төмен болса, жақсырақ, ансамбль техникасында да кішірек болды (0,30375 қарсы 0,46008), бұл кластердің ықшамдылығын және аз қабаттасуды бейнелейді. Бұл өлшемдер базалық KMeans алгоритмімен салыстырғанда ансамбль моделінің жақсырақ кластерлеуге қол жеткізетінін растайды.



Сур. 4. Ішкі бағалау метрикасы.

5-суретте екі тәсіл үшін сыртқы кластерлеу сапасының көрсеткіштері салыстырылады: KMeans алгоритміне негізделген қарапайым үлгі (қызғылт сары жолақтар) және біріктірілген кластерлеу тәсілін пайдаланатын ансамбльдік үлгі (көк жолақтар). Бағалау көрсеткіштері ретінде Adjusted Rand Index (ARI), Normized Mutual Information (NMI) және детерминация коэффициенті (R^2 Score) пайдаланылды. Барлық көрсеткіштер ансамбль үлгісінің маңызды артықшылығын көрсетеді. Осылайша, ARI және NMI мәндері ансамбль үшін 0,9996 болса, KMeans үшін 0,97879 болды, бұл тиісті белгілермен жоғары сәйкестікті және кластерлеудің үлкен ақпараттандырғыштығын көрсетеді. Сол сияқты, R^2 Score мәні — ансамбль үшін 0,99889 және KMeans үшін 0,976 — үлгінің жақсырақ түсіндіру күшін көрсетеді. Осылайша, ансамбльдік тәсіл барлық үш көрсеткіш растайтын нақты деректер құрылымына кластерлердің дәлірек сәйкестігін қамтамасыз етеді.



Сур. 5. Сыртқы бағалау метрикасы.

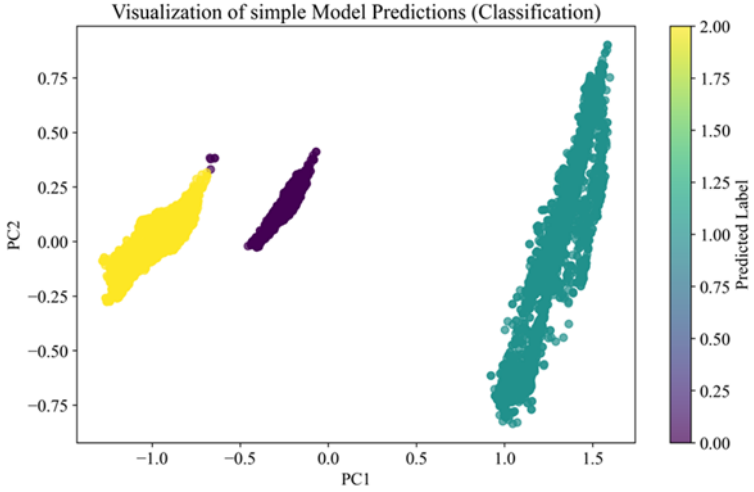
6-суретте екі тәсіл үшін кластерлеу қателері салыстырылады: KMeans алгоритмін пайдаланатын негізгі модель (қызғылт сары жолақтар) және бірнеше әдістерді біріктіретін ансамбльдік модель (көк жолақтар). Өлшем ретінде орташа квадрат қатесі (RMSE) және орташа абсолютті қателік (MAE) қолданылды, бұл бізге болжамды және анықтамалық мәндер арасындағы айырмашылықты салыстыруға мүмкіндік берді. Ансамбльдік модельдің қателік мәндерінің едәуір аз екенін байқауға болады: MAE шамамен 0,00005-ке қарсы 0,00009-ға қарсы қарапайым модель үшін, ал RMSE шамамен 0,00007-ге қарсы 0,0001. Бұл дәлірек болжауды және модельдің нақты деректерді таратуға жақынырақ жақындауын көрсетеді. Осылайша, ансамбльдік тәсіл құрылымдық тұрғыдан анық кластерлеуді қамтамасыз етеді және сандық қателерді азайтады, бұл модельді регрессиялық талдау мәселелеріне немесе топырақтың тұздану дәрежесі сияқты ерекшеліктерді сандық бағалауға қолдану кезінде әсіресе маңызды.

Сур. 6. Әртүрлі кластерлеу әдістері үшін MAE және RMSE талдауы.

Ансамбль үлгісінің жоғары Silhouette Score мәні (қарапайым KMeans үлгісі үшін 0,76754 және 0,61703) жақсы кластерлеу сапасын көрсетеді — кластердегі нысандар ықшамырақ орналасады және кластерлер арасындағы қашықтық ұлғаяды. Бұл ұсынылған алгоритм деректерді бөлуді жақсырақ жобалайтынын білдіреді: бірдей кластері бар нысандар айтарлықтай ұқсастыққа ие, ал әртүрлі кластерлер арасындағы айырмашылықтар жақсарады. Сонымен қатар, KMeans бойынша Дэвис-Боулдин ұпайының 0,46008 ұпайынан ансамбльдік шешімнің 0,30375-ке дейін төмендеуі кластерлер арасындағы тығыз қабаттасуды және көп өлшемді және шулы деректер жиынын қарастырған кезде әсіресе қажет болатын кластер ішілік біртектіліктің жоғарылауын көрсетеді. Сыртқы көрсеткіштер де кластерлеу сапасын жақсартады. Ансамбль үлгісінің Adjusted Rand Index (ARI) және Norized Mutual Information (NMI) мәндері 0,9996 болды, ал қарапайым үлгінікі 0,97879 болды, бұл болжанған кластерлер мен нақты белгілер арасындағы жоғары сәйкестікті көрсетті. Сол сияқты, R^2 ұпайы KMeans үшін 0,976-дан ансамбль үшін 0,99889-ға дейін өсті, бұл модельдің жақсырақ түсіндіру күшін көрсетеді. Ансамбльдік модель қателер бойынша да жақсырақ мәндерді көрсетті: MAE қарапайым модель үшін 0,00009-дан 0,00005-ке дейін және RMSE 0,0001-ден 0,00007-ге дейін төмендеді. Бұл дәлірек сандық болжамдарды және анықтамалық мәндерден аз ауытқуды көрсетеді. Ғылыми тұрғыдан алғанда, барлық көрсеткіштердің жетілдірілуі ұсынылып отырған тәсілдің бір ғана кластерлеу алгоритмімен шектелмейтіндігінде. KMeans псевдобелгілерінде оқытылған XGBoost пайдалану деректердегі жасырын үлгілерді анықтауға және бастапқы мүмкіндік кеңістігін ықтималдық сипаттамалармен байытуға, оның ақпараттық мазмұнын арттыруға мүмкіндік береді. KMeans, агломеративті кластерлеу және DBSCAN біріктіретін ансамбльдік механизм деректер құрылымының әртүрлі көріністерін қарастыруға мүмкіндік береді, ал нәтижелер бойынша

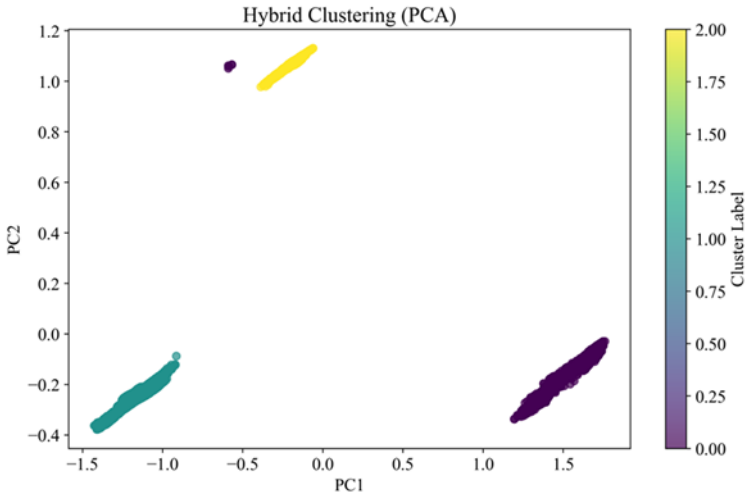
дауыс беру жеке алгоритмдердің кемшіліктерімен байланысты қателердің ықтималдығын азайтады. Кластерлеу кезеңінде алынған белгілердің жоғары сенімділігіне байланысты кейбір объектілерді кейінгі бақылаудағы оқыту үшін пайдалануға болады. Бұл классификациялық есептерді (топырақ түрін анықтау) және регрессияны (тұздану дәрежесін сандық бағалау) бір мезгілде шешетін модель құруға мүмкіндік береді. Осылайша, ұсынылған әдіс деректер құрылымын ерекшелейтін бақыланбайтын әдістердің де, дәлірек, сенімді және түсіндірілетін болжамдарды қамтамасыз ететін сыныптардың шекараларын нақтылайтын бақыланатын әдістердің де күшті жақтарын біріктіреді.

7-суретте PCA арқылы алынған екі негізгі құрамдасқа (PC1 және PC2) қатысты деректер көрсетілген. Нүктелер қарапайым үлгінің болжамды белгілеріне сәйкес боялады, мұнда түс айырмашылықтары әртүрлі сыныптарды көрсетеді. Объектілердің бірнеше сегменттерге топтастырылғанын көруге болады, бірақ кейбір нүктелер арасында кейбір қабаттасулар бар және барлық топтардың нақты белгіленген шекаралары жоқ. Бұл модельдің жалпы бөлуді жасайтынын көрсетеді, бірақ ең алдымен спектрлік кеңістіктегі деректер қабаттасатын немесе жеткілікті түрде дифференциацияланбаған болса, шекті қателер болуы мүмкін.



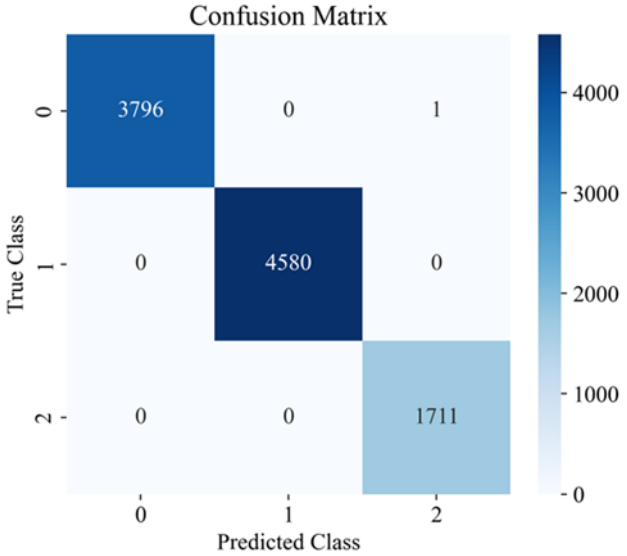
Сур. 7. Қарапайым Үлгі Болжамдарының визуализациясы (Классификация).

8-суретте кластерлеу нәтижелері көрсетілген. Ол сонымен қатар екі өлшемді деректерді проекциялау үшін негізгі құрамдастарды пайдаланады. Дегенмен, соңғы белгілер бірнеше кластерлеу әдістерін (KMeans, агломеративті кластерлеу және DBSCAN) біріктіру және XGBoost арқылы мүмкіндіктерді байыту арқылы алынды. График объектілердің топтарға неғұрлым нақты бөлінуін көрсетеді: кластерлер олардың арасында байқалатын бос орындармен бөлек орналасқан. Бұл ұсынылған тәсіл деректердегі көп өлшемді тәуелділіктерді жақсырақ есептей алатынын көрсетеді, бұл дәлірек және сенімді сегментацияға әкелді.



Сур. 8. Гибридті кластерлеу (PCA).

Ансамбль үлгісі кластерлеу сапасы мен бөлу айқындылығында сенімдірек нәтиже көрсетеді. Ол жақсы ажыратылған кластерлерді қалыптастырып қана қоймайды, сонымен қатар жіктелуі қиын шекаралық нысандардың азырақ санына ие. Бұл ұсынылған тәсіл үшін жоғары Silhouette мәндерін және төменірек Дэвис-Боулдин мәнін көрсететін сандық көрсеткіштермен де расталады. 9-суретте болжамды класс көлденеңінен, ал шынайы класс тігінен салынған. Ұяшықтардың ішіндегі сандар нақты болжанған сыныпқа үлгі тағайындалған әрбір нақты сыныптың (0,1 немесе 2) қанша нысанын көрсетеді. Барлық дерлік нысандар диагональда шоғырланғандығы анық: 0-сыныптан 3796-сы 0,1-сыныптан 4580-і 1, ал 2-сыныптан 1711-і 2 деп дұрыс болжалды. Жалғыз қателік 2-сыныпқа қате тағайындалған (бірінші жолдағы ұяшық, үшінші бағандағы ұяшық) нақты 0-сыныптың бір объектісі болып табылады.



Сур. 9. Классификация моделінің шатастыру матрицасы өнімділік

Осылайша, ұсынылған нәтижелер ұсынылған модель барлық негізгі көрсеткіштер бойынша KMeans алгоритміне негізделген базалық модельден асып, кеңістіктік сегменттеуді және топырақтың тұздылығын сандық бағалауды тиімді жеңетінін көрсетеді. Sil-houette Score және ARI жоғары мәндері, Дэвис-Боулдин ұпайының төмендеуі және MAE және RMSE қателерінің минималды мәндері кластерлердің құрылымдық анықтығын ғана емес, сонымен қатар сандық болжамдардың жоғары дәлдігін растайды. PCA көмегімен деректерді визуализациялау ансамбль үлгісінің сандық көрсеткіштерге сәйкес келетін оқшауланған және түсіндірілетін топтарды құрайтынын көрсетті. Жіктеу қателерінің ең аз саны (10 мыңнан астам жазбаның біреуі ғана) анықталған сыныптардың жоғары сенімділігін көрсетеді. Мұның бәрі бақыланбайтын әдістердің, ықтималды байытудың және нейрондық желіні көп тапсырмалы оқытудың кешенді комбинациясы дәстүрлі тәсілдерге қарағанда дәлірек, сенімді және практикалық нәтижелер беретінін растайды.

Қорытынды

Бұл зерттеу Sentinel-2 спутниктік деректері мен ERA5-Land климаттық параметрлерін пайдалана отырып, топырақтың тұздылығын болжау және сегменттеу мәселесіне кешенді тәсілдің тиімділігін көрсетті. Әзірленген модель бақыланбайтын әдістерді (KMeans, Агломеративті кластерлеу, DBSCAN), бақыланатын оқытуды (XGBoost) және гетерогенді кеңістіктік деректермен жұмыс істеу кезінде жоғары дәлдік пен сенімділікке қол жеткізген көп тапсырмалы нейрондық желіні біріктіреді. Негізгі кластерлік көрсеткіштер сапаның жақсарғанын көрсетті: Силуэт ұпайы 0,8156-ға жетті, ал Дэвис-Боулдин ұпайы 0,3022-ге дейін төмендеді. Классификациялық тапсырмада модельдің дәлдігі 99,99 % құрады, 10 000-нан астам сынақ бақылауларында бір рет қате жіберілді.

Ұсынылған әдіс топырақтың әртүрлі типтерінің спектрлік ерекшеліктерін бейімдеуге, шекаралық қателіктерді азайтуға және сортаңдану процестеріне климаттық факторлардың әсерін ескеруге мүмкіндік береді. Ансамбльдік әдістер мен мүмкіндіктерді ықтималдықпен байыту арқасында дәстүрлі алгоритмдерді пайдаланғаннан гөрі топырақ кластарын дәлірек бөлуге қол жеткізу мүмкін болды.

Алынған нәтижелер ұсынылған тәсілдің ғылыми және қолданбалы маңыздылығын растайды. Оны жердің тозуын бақылау, құнарлылықтың төмендеуі қаупін бағалау және агроэкологиялық жүйелерде шешім қабылдауды қолдау үшін пайдалануға болады. Әдістемені конволюционды нейрондық желілерді қолдану арқылы терең сегменттеу бағытында дамыту, экипажсыз әуе кемелерінің мәліметтерін біріктіру және талдауды басқа аймақтарға және маусымдық интервалдарға кеңейту жоспарлануда. Бұл климаттың өзгеруі жағдайында топырақтың тұзденуін бақылаудың әмбебап және ауқымды жүйесін қамтамасыз етеді.

REFERENCES

Aksoy S., Sertel E., Roscher R., Tanik A., Hamzehpour N. (2024). Assessment of soil salinity using explainable machine learning methods and Landsat 8 images // International Journal of Applied Earth Observation and Geoinformation. — 2024. — Vol. 130. Article 103879. DOI: 10.1016/j.jag.2023.103879.

Becker S. J., Maloney M. C., Griffin A. W., Lasko K., Sussman H. S. (2025). Bare ground classification using a spectral index ensemble and machine learning models optimized across 12 international study sites // Geocarto International. — 2025. — Vol. 40. — No. 1. Article 2465452. DOI: 10.1080/10106049.2023.2465452.

Bo Q., Lv P., Wang Z., Wang Q., Li Z. (2024). Prediction of the post-mining land use based on random forest and DBSCAN // PLoS ONE. — 2024. — Vol. 19. — No. 1. Article e0287079. DOI: 10.1371/journal.pone.0287079.

Bougiouklis J. N., Barouchas P. E., Petropoulos P., Tsesmelis D. E., Moustakas N. (2025). Precision soil sampling strategy for the delineation of management zones in olive cultivation using unsupervised machine learning methods // Scientific Reports. — 2025. — Vol. 15. — No. 1. Article 8253. DOI: 10.1038/s41598-024-41188-9.

H. Tavakoli Computers and Electronics in Agriculture (2024). Integrating satellite and ground-based measurements for predicting soil salinity // Computers and Electronics in Agriculture. — 2024. (Article details not provided).

Computers and Electronics in Agriculture (2025). Automated soil salinity detection using Sentinel-2 data and deep learning models // Computers and Electronics in Agriculture. — 2025. (Article details not provided).

Cui G., Liu Y., Li X., Wang S., Qu X., Wang L., Zhang W. (2025). Impacts of groundwater storage variability on soil salinization in a semi-arid agricultural plain // Geoderma. — 2025. — Vol. 454. Article 117162.

Environmental Challenges (2024). UAV-based mapping of soil salinity using geostatistical and deep learning approaches // Environmental Challenges. — 2024. (Article details not provided).

Gao L., Xu E. (2025). Mapping cropland soil salinity using multi-cycle classification to mitigate retrieval errors // Computers and Electronics in Agriculture. — 2025. — Vol. 231. Article 110055. DOI: 10.1016/j.compag.2024.110055.

Gao Z., Peña-Arancibia J. L., Siyal A. A. (2025). Three-dimensional soil salinity mapping with uncertainty using Bayesian hierarchical modelling, random forest regression and remote sensing data // Agricultural Water Management. — 2025. — Vol. 309. Article 109318. DOI: 10.1016/j.agwat.2024.109318.

Geoderma (2024). Machine learning-based predictions of soil degradation using ERA5-Land data // Geoderma. — 2024. (Article details not provided).

Jia P., Zhang J., Liang Y., Zhang S., Jia K., Zhao X. (2024). The inversion of arid-coastal cultivated soil salinity using explainable machine learning and Sentinel-2 // Ecological Indicators. — 2024. — Vol. 166. Article 112364. DOI: 10.1016/j.ecolind.2023.112364.

Jiang Z., Ding J., Li Z., Liu J. (2025). Present knowledge and future challenges in remote sensing for soil salinization monitoring: A review of bibliometric analysis // International Journal of Remote Sensing. — 2025. — Vol. 46. — No. 1. Pp. 247–272. DOI: 10.1080/01431161.2024.2412804.

Kaliyeva D., Tokbergenova A., Mirzabaev A., Zulpykharov K., Bissenbayeva S., Taukebayev O., Qadir M. (2025). Assessing soil erosion risk in Kazakhstan: A RUSLE-based approach for land rehabilitation // Polish Journal of Environmental Studies. — 2025. — Vol. 34. — No. 3.

Land (2024). Multispectral remote sensing for soil salinity assessment in arid regions // Land. — 2024. (Article details not provided).

Liu X., Hu Y., Li X., Du R., Xiang Y., Zhang F. (2024). An interpretable model for salinity inversion assessment of the South Bank of the Yellow River based on Optuna hyperparameter optimization and XGBoost // Agronomy. — 2024. — Vol. 15. — No. 1. Article 18. DOI: 10.3390/agronomy15010018.

Luo Z., Deng M., Tang M., Liu R., Feng S., Zhang C., Zheng Z. (2025). Estimating soil profile salinity under vegetation cover based on UAV multi-source remote sensing // Scientific Reports. — 2025. — Vol. 15. — No. 1. Article 2713. DOI: 10.1038/s41598-024-40157-9.

Lusiana E. D., Astutik S., Nurjannah N., Sambah A. B. (2025). Using machine learning approach to cluster marine environmental features of Lesser Sunda Island // Journal of Applied Data Science. — 2025. — Vol. 6. — No. 1. Pp. 247–258. DOI: 10.1007/s42488-024-00127-1.

Lusiana Vediappan D., Godi A., Golla B. (2025). Geospatial mapping of soil properties of forest types using the k-means fuzzy clustering approach to delineate site-specific nutrient management zones in Goa, India // Journal of the Indian Society of Remote Sensing. — 2025. (Volume and pages not provided). DOI: 10.1007/s12524-024-01813-5.

Paramonova T.A., Shynbergenov Y.A., Botavin D.V., Golosov V.N. (2025). Assessment of soil pollution and

erosion processes in the Republic of Kazakhstan according to literature data // Eurasian Soil Science. — 2025. — Vol. 58. — No. 1. P. 11. DOI: 10.1134/S1064229325010083.

Sadyrova G., Tynybekov B., Nazarbekova S., Orazbekova K., Ibragimov T., Shimshikov B., Nurmakhanova A. (2025). Impact of cattle grazing on degradation of mountain pastures in South-East Kazakhstan // ES Energy & Environment. — 2025. (Article details not provided).

Tashpolat N., Reheman A. (2025). Monitoring of soil salinity in the Weiku Oasis based on feature space models with typical parameters derived from Sentinel-2 MSI images // Land. — 2025. — Vol. 14. — No. 2. Article 251. DOI: 10.3390/land14020251.

Tussupov J., Yessenova M., Abdikerimova G., Aimbetov A., Baktybekov K., Murzabekova G., Aitimova U. (2024). Analysis of formal concepts for verification of pests and diseases of crops using machine learning methods // IEEE Access. — 2024. — Vol. 12. Pp. 19902–19910. DOI: 10.1109/ACCESS.2024.3363886.

Tussupov J., Abdikerimova G., Ismailova A., Kassymova A., Beldeubayeva Z., Aitimov M., Makulov K. (2024). Analyzing disease and pest dynamics in steppe crop using structured data // IEEE Access. — 2024. — Vol. 12.

Velilla E., Snethlage J., Poelman M., van der Meer I.M., van der Werf A., Deolu-Ajayi A.O., van Belzen J. (2025). Too salty to farm: Rethinking coastal land use in response to soil salinization // Restoration Ecology. — 2025. — Vol. 33. Article e70006. DOI: 10.1111/rec.70006.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 43–58
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.003

ISSUES OF HANDOVER IN ULTRA-DENSE NETWORKS

A.S. Abdiraman[1], L.S. Aldasheva[1], D. Sagidullauly[1], A.S. Amirova[1], A. Tanirbergenova[2]

[1]Astana IT University, Astana, Kazakhstan;
[2]Euroasian National University named after L.N. Gumilyev, Astana, Kazakhstan.
E-mail: a.s.abdiraman@gmail.com

Aliya Abdiraman — master, senior lecturer of the Cybersecurity School, Astana IT University, Astana, Kazakhstan
E-mail: a.s.abdiraman@gmail.com, https://orcid.org/0000-0001-5494-0223;
Laura Aldasheva — c.t.s., assistant professor of the Cybersecurity School, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-6815-1989;
Dauren Sagidullauly — master, teacher of the Cybersecurity School, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0004-0996-0275;
Akzhibek Amirova — PhD, assistant professor of the Cybersecurity School, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-5715-4954;
Alua Tanirbergenova — PhD, senior lecturer of the Cybersecurity School, Euroasian National University named after L.N. Gumilyev, Astana, Kazakhstan
https://orcid.org/0000-0001-8555-7315.

Abstract. This article analyzes current research in the field of mobility management in ultra-dense networks. It examines key challenges and promising approaches to ensuring stable connectivity in high-density device environments with small cell sizes. The transition to 5G/6G networks, the implementation of the Internet of Things (IoT), the development of drone-based networks, and the use of millimeter waves impose new requirements for handover management and network resource adaptation. Particular attention is given to the impact of new communication standards on ICT infrastructure architecture, including the integration of wireless networks with cloud computing, quantum communications, and artificial intelligence. Both traditional mobility management algorithms and adaptive mechanisms based on machine learning, self-optimization, and intelligent systems are considered. The article provides

an overview of modern methods for ensuring connection continuity in high-mobility networks, analyzing their efficiency and limitations. Additionally, it discusses future directions for ICT infrastructure development aimed at improving data transmission reliability and reducing latency in next-generation ultra-dense networks.

Keywords: 5G, 6G, ultra-dense networks, handover, mobility management, artificial intelligence

For citation: A.S. Abdiraman, L.S. Aldasheva, A.S. Amirova, D. Sagidullauly, A.S. Tanirbergenova. Issues of handover in ultra-dense networks // International jour-nal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 43–58. (In Russ.). https://doi.org/10.54309/IJICT.2025.24.4.003.

Conflict of interest: The authors declare that there is no conflict of interest.

АСА ТЫҒЫЗЖЕЛІЛЕРДЕГІ ҚОСЫЛУДЫ ТАСЫМАЛУ МӘСЕЛЕЛЕРІ

Ә.С. Әбдіраман[1], Л.С. Алдашева[1], А.С. Амирова1, Д. Сағидуллұлы[1], А. Танирбергенова[2]

[1]Astana IT University, Астана, Қазақстан;
[2]Л.Н. Гумилев атындағы ЕҰУ, Астана, Қазақстан.
E-mail: a.s.abdiraman@gmail.com

Әбдіраман Әлия — Магистр, Киберқауіпсіздік Мектебінің сеньор лекторы, Astana IT University
E-mail: a.s.abdiraman@gmail.com. https://orcid.org/0000-0001-5494-0223;
Алдашева Лаура — т.ғ.к, Киберқауіпсіздік Мектебінің ассистент профессоры, Astana IT University
https://orcid.org/0000-0001-6815-1989;
Сағидуллаұлы Даурен — Киберқауіпсіздік Мектебінің сеньор лекторы, Astana IT University
https://orcid.org/0009-0004-0996-0275;
Амирова Акжибек — PhD, Киберқауіпсіздік Мектебінің ассистент профессоры, Astana IT University
https://orcid.org/0000-0002-5715-4954;
Танирбергенова Алуа — PhD, Жүйелік анализ және басқару кафедрасының аға оқытушысы, Л.Н. Гумилев атындағы ЕҰУ
https://orcid.org/0000-0001-8555-7315.

Аннотация. Бұл мақалада аса тығыз желілердегі мобильділікті басқару саласындағы өзекті зерттеулерге талдау жасалған. Жоғары құрылғылар тығыздығы мен шағын ұяшық өлшемдері жағдайында тұрақты қосылымды қамтамасыз етудің негізгі мәселелері мен перспективалық

тәсілдері қарастырылады. 5G/6G желілеріне көшу, заттар интернетін (IoT) енгізу, дрон желілерін дамыту және миллиметрлік толқындарды пайдалану хэндоверді басқару мен желілік ресурстарды бейімдеуге жаңа талаптар қояды. Әсіресе, жаңа байланыс стандарттарының бұлттық есептеулермен, кванттық коммуникациялармен және жасанды интеллектпен біріктірілген сымсыз желілерге әсеріне ерекше назар аударылады. Мақалада дәстүрлі мобильділікті басқару алгоритмдерімен қатар, машиналық оқытуға, өзін-өзі оңтайландыруға және интеллектуалды жүйелерге негізделген бейімделгіш механизмдер қарастырылады. Жоғары мобильділік желілерінде қосылымның үздіксіздігін қамтамасыз етудің заманауи әдістеріне шолу жасалып, олардың тиімділігі мен шектеулері талданады. Сонымен қатар, жаңа буын аса тығыз желілерінде деректерді берудің сенімділігін арттыруға және кідірісті азайтуға бағытталған ИКТ инфрақұрылымын дамытудың болашақ бағыттары талқыланады.

Түйін сөздер: 5G, 6G, өте тығыз желілер, тапсыру, ұтқырлықты басқару, жасанды интеллект

Дәйексөздер үшін: Ә.С. Әбдіраман, Л.С. Алдашева, А.С.Амирова, Д.Сагидуллаулы, А.Танирбергенова. Аса тығызжелілердегі қосылуды тасымалу мәселелері // еждународный журнал информации и коммуникационной тех-нологии. 2025. Vol. 6. No. 24. Pp. 43–58. (Орыс тіл.). https://doi.org/10.54309/IJICT.2025.24.4.003.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ПРОБЛЕМЫ ПЕРЕДАЧИ СОЕДИНЕНИЯ В СВЕРХПЛОТНЫХ СЕТЯХ

А.С. Абдираман[1], Л.С. Алдашева[1], А.С. Амирова[1], А. Танирбергенова[2], Д. Сагидуллаулы[1]

[1]Астана ИТ университет, Астана, Казахстан;
[2]ЕНУ им. Л.Н. Гумилева, Астана, Казахстан.
E-mail: a.s.abdiraman@gmail.com

Абдираман Алия — магистр, сениор лектор Департамента интеллектуальных систем и кибербезопасности, Астана ИТ университет
E-mail: a.s.abdiraman@gmail.com, https://orcid.org/0000-0001-5494-0223;
Алдашева Лаура — к.т.н., ассистент профессор Школы кибербезопасности, Астана ИТ университет
https://orcid.org/0000-0001-6815-1989;
Сагидуллаулы Даурен — магистр, преподаватель Школы кибербезопасности, Астана ИТ университет
https://orcid.org/0009-0004-0996-0275;
Амирова Акжибек — PhD, ассистент профессор Школы кибербезопасности, Астана ИТ университет

https://orcid.org/0000-0002-5715-4954;
Танирбергенова Алуа — PhD, старший преподаватель, кафедры системного анализа и управления ЕНУ им. Л.Н. Гумилева
https://orcid.org/0000-0001-8555-7315.

Аннотация. В данной статье проведен анализ актуальных исследований в области управления мобильностью в сверхплотных сетях. Рассматриваются основные проблемы и перспективные подходы к обеспечению стабильного соединения в условиях высокой плотности устройств и малых размеров сот. Переход к сетям 5G/6G, внедрение интернета вещей (IoT), развитие дроновых сетей и использование миллиметровых волн предъявляют новые требования к управлению хэндовером и адаптации сетевых ресурсов. Особое внимание уделяется влиянию новых стандартов связи на архитектуру ИКТ-инфраструктуры, включая интеграцию беспроводных сетей с облачными вычислениями, квантовыми коммуникациями и искусственным интеллектом. Рассматриваются как традиционные алгоритмы управления мобильностью, так и адаптивные механизмы, основанные на машинном обучении, самооптимизации и интеллектуальных системах. В статье приведен обзор современных методов обеспечения непрерывности соединения в сетях с высокой мобильностью, анализируются их эффективность и ограничения. Также обсуждаются перспективные направления развития ИКТ-инфраструктуры, направленные на повышение надежности передачи данных и снижение задержек в условиях сверхплотных сетей нового поколения.

Ключевые слова: 5G, 6G, сверхплотные сети, хэндовер, управление мобильностью, искусственный интеллект, ИКТ

Для цитирования: А.С. Абдираман, Л.С. Алдашева, А.С. Амирова, Д. Сагидуллаулы, А. Танирбергенова. Проблемы передачи соединения в сверхплотных сетях // Международный журнал информации и коммуникационной технологии. 2025. Vol. 6. No. 24. Pp. 43–58. (На русс.). https://doi.org/10.54309/IJICT.2025.24.4.003.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

За последние три десятилетия мобильные сети претерпели значительное революционное развитие (Kamel et al., 2016). Потенциал 5G и 6G значительно расширился, что привело к разработке множества новых сценариев использования и применению принципиально новых эффективных подходов. Постоянное развитие систем мобильной связи 5G и 6G выявляет существенные ограничения этих сетей. Мобильная сеть 5G уже развернута в некоторых регионах мира.

Ожидается, что к концу 2025 года около 65% населения мира будет иметь доступ к сети 5G (Chen et al., 2016), а к 2026 году их число достигнет 3,5 миллиарда (Hao et al., 2016).

В сетях 5G традиционно выделяют три ключевых усовершенствования: расширенный мобильный широкополосный доступ (enhanced mobile broadband, eMBB), массовые коммуникации между устройствами (massive machine-type communications, mMTC) и ультранадежные связи с низкой задержкой (ultra-reliable low-latency communications, URLLC). Каждое из этих направлений имеет высокие требования, такие как обеспечение более широкой зоны покрытия, увеличение емкости сети, высокая надежность и минимизация задержек. Очевидно, что каждый сценарий использования 5G требует различных стратегий передачи данных, влияющих на нагрузку сигнального трафика, энергопотребление и задержку передачи (Banafaa et al., 2024).

Кроме того, 5G обеспечивает широкие возможности подключения, включая Интернет вещей (IoT), взаимодействие «машина-машина» (M2M), связь «устройство-устройство» (D2D), технологии «автомобиль-все» (V2X) и Bluetooth. В совокупности эти технологии повлияют на способы взаимодействия бизнеса, государственных структур и потребителей в физическом мире (Gures et al., 2020). В течение следующего десятилетия указанные выше сервисы станут ключевыми компонентами крупнейших мировых рынков устройств (Alraih et al., 2022).

Следовательно, ожидается, что в сетях 5G появятся сотни тысяч одновременных подключений, необходимых для массового развертывания данных сервисов, а в предстоящих сетях 6G эта задача станет еще более сложной (Покаместов et al., 2024). Эти разнообразные типы подключенных сервисов предъявляют высокие требования к скорости передачи данных и требуют чрезвычайно широкого спектра частот.

Материалы и методы.

Эволюция и перспективы развития мобильных сетей: от 5G к 6G

Разработка и внедрение сетей 5G (Fifth Generation — пятое поколение беспроводной мобильной связи) неизбежно привело к возникновению новых концепций построения сетей мобильной связи (Angjo et al., 2021). Огромный объем трафика, обусловленный потребностью современного поколения, требует изменения парадигмы во всех аспектах работы мобильных сетей. Новые сетевые сервисы, такие как видео высокого разрешения, дополненная реальность, облачные вычисления и интернет вещей определяют требования к современным сетям. Сети 5G позволяют в сотни раз увеличить пропускную способность по сравнению с предыдущим поколением 4G (Fourth Generation — четвертое поколение беспроводной мобильной связи). Развитие концепции было продиктовано внедрением поддержки сот малых размеров (фемтосоты, пикосоты и микросоты). Несмотря на статус внедренного стандарта 5G, исследования в области многих аспектов в рамках стандарта продолжаются до сих пор (Shayea

et al., 2022).

В настоящее время ведутся активные исследования по разработке будущего стандарта мобильной связи 6G (Sixth Generation — шестое поколение беспроводной мобильной связи). Технология шестого поколения (6G) мобильных сетей установит новые стандарты производительности мобильных сетей (Alshaibani et al., 2022). Это связано с высокими требованиями к новым сервисам на основе искусственного интеллекта, сверхнизким задержкам, экстремальной скорости передачи данных в сети и поддержке огромного количества различных подключенных приложений. Высокие требования к скорости передачи и задержке вынуждают искать решение в новых частотных диапазонах (рис. 1). Возросшая сложность будущего стандарт мобильной связи привлекает внимание огромного количества исследователей со всего мира, среди которых лидируют коллективы из КНР, Южной Кореи, США и Европы.

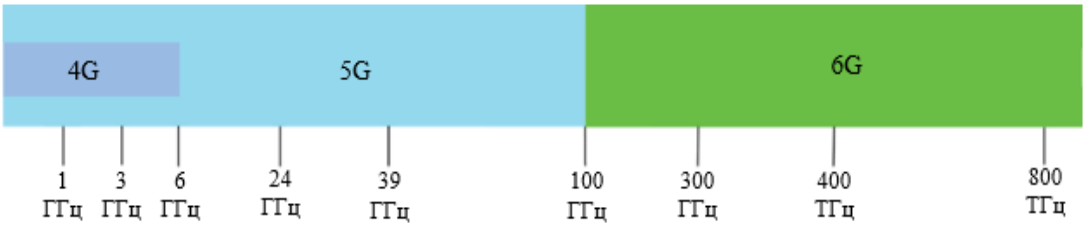


Рис.1. Частотный диапазон 4G/5G и ожидаемый диапазон 6G

Сверхплотный сценарий развертывания сети будет ключевым для стандарта 6G, но он также применим и для сетей пятого поколения. Несмотря на концентрацию особого внимания вокруг активной разработки стандарта 6G, исследования в области 5G продолжаются, поэтому в данной работе представлен обзор публикаций в рамках обоих стандартов связи. Развертывание сверхплотных базовых станций на малых сотах в мобильных сетях следующего поколения – одна из важнейших задач для исследователей. Управление мобильностью – критический вопрос, требующий особого внимания для достижения высоконадежного соединения мобильных пользовательских устройств. Задача усложняется тем, что новый стандарт мобильной связи должен быть полностью совместим с предыдущим поколением, например, при внедрении сверхплотных малых сот 5G параллельно с текущими сетями 4G. Таким образом, исследования в области управления хэндовером в будущих сценариях развертывания мобильных сетей в настоящий момент являются очень востребованными и актуальными.

В таблице 1 представлено сравнение ключевых характеристик стандартов 4G, 5G и 6G, включая частотный диапазон, задержку, скорость передачи данных и поддержку устройств. Эти параметры демонстрируют эволюцию мобильных сетей от четвертого к шестому поколению и отражают их влияние на современные технологии и приложения.

Таблица 1. Сравнение ключевых характеристик стандартов 4G, 5G и 6G

Характеристика	4G			5G	6G	
Частотный диапазон		2-8 ГГц		3-100 ГГц		100 ГГц – 1ТГц
Задержка	50 мс	1–10 мс	Менее 1 мс			
Скорость передачи данных		До 1 Гбит/с		До 20 Гбит/с		До 1 Тбит/с
Поддержка устройств	Ограниченная поддержка IoT, базовые устройства	Массовая поддержка IoT, смартфоны, дроны	Интеллектуальные устройства, IoT 2.0, квантовые сети			

В представленной таблице показано сравнение ключевых характеристик стандартов мобильной связи 4G, 5G и 6G, демонстрирующее эволюцию технологий. Частотный диапазон увеличился с 2-8 ГГц в 4G до 3-100 ГГц в 5G и до ожидаемого диапазона 100 ГГц - 1 ТГц в 6G. Такое расширение позволяет использовать более высокие частоты для передачи данных, что способствует увеличению пропускной способности. Задержка соединения значительно сократилась: с 50 мс в 4G до 1–10 мс в 5G и менее 1 мс в 6G, что делает сети следующего поколения более подходящими для приложений, требующих мгновенного отклика.

Таблица также демонстрирует экспоненциальный рост скорости передачи данных: от 1 Гбит/с в 4G до 20 Гбит/с в 5G и ожидаемого 1 Тбит/с в 6G. Что касается поддержки устройств, 4G предоставляет ограниченные возможности для IoT, тогда как 5G массово поддерживает IoT, смартфоны и дроны. В 6G акцент будет сделан на интеллектуальных устройствах, IoT 2.0 и квантовых сетях, что откроет новые горизонты для технологий умных городов, промышленной автоматизации и здравоохранения. Эта таблица подчеркивает ключевые изменения, определяющие будущее мобильных сетей.

Согласно терминологии мобильных сетей, хэндовером называют процесс переключения мобильной станции от одной базовой станции к другой. Сбой хэндовера является одной из основных проблем мобильности, поэтому одной из основных задач 5G является управление мобильностью. Развертывание большого количества малых сот в сети приводит к увеличению границ сот и усилению межсотового взаимодействия, что приводит к частым переключениям между базовыми станциями и увеличению числа отказов. Негативное частое переключение абонентского терминала между базовыми станциями в англоязычной литературе называется пинг-понгом хэндовера, или PPHO (Ping-Pong Handover). PPHO снижает производительность сети, поскольку приводит к излишнему потреблению энергии, что критично для мобильных

устройств. Дополнительная нагрузка на сеть возникает при отказах хэндовера, поэтому в рамках исследований в области управления мобильностью данные факторы должны учитываться в первую очередь.

Результаты и обсуждение.

Влияние 5G/6G в развитие ИКТ

Поддержка бесперебойного соединения мобильных пользователей – важнейшая задача, которая требует решения в мобильных гетерогенных сетях (HetNet). Проблемы обусловлены различными факторами, такими как использование миллиметровых волн, массовый рост подключенного мобильного оборудования, дублирование развертывания сетей, внедрение двойного подключения, агрегация частот, появление подключенных БПЛА, массовое развертывание малых сот и требования к скорости передачи. В работе (Khan et al., 2022) представлен всесторонний обзор управления хэндовером в будущих мобильных сверхплотных сетях HetNet. В ней обсуждаются различные методы управления мобильностью и хэндовером, рассматриваются некоторые ключевые проблемы, возможности и перспективные решения.

Авторы (Alhammadi et al., 2018) предложили динамическое управление параметрами хэндовера в сетях HetNet с плотными малыми ячейками. Результаты моделирования показывают, что предложенный алгоритм значительно снижает вероятность скачкообразных переключений и прерывания связи, тем самым улучшая производительность сети.

В работе (Shayea et al., 2020) рассматриваются ключевые факторы, которые будут в значительной степени способствовать росту проблем мобильности. Кроме того, обсуждаются инновационные, передовые, эффективные и интеллектуальные методы передачи данных, которые были внедрены в сетях 5G. В исследовании также выделены основные проблемы существующих сетей, а также направления будущих исследований в области управления мобильностью в сетях 5G и выше.

Авторы (Alhammadi et al., 2020) предлагают алгоритм самооптимизации на основе скорости для управления параметрами хэндовера в сетях 4G/5G. Предложенный алгоритм использует полученную мощность и скорость пользователя для настройки запаса HO и времени срабатывания во время мобильности пользователя в сети. Результаты моделирования показывают, что предложенный алгоритм позволяет значительно сократить количество скачков передачи при хэндовере по сравнению с другими существующими алгоритмами, превосходя их в среднем более чем на 70 % по всем показателям производительности HO.

В исследовании (Alraih et al., 2022) предлагаются различные настройки системы управления хэндовером для изучения и анализа в сетях 5G и выше. Результаты показывают, что различные настройки системы оказывают различное и значительное влияние на производительность сети. В работе продемонстрировано, что для повышения эффективности управления

хэндовером параметры системы должны быть тщательно подобраны с учетом мобильной среды и сценария использования.

Одна из важнейших технологий хэндовера известна как оптимизация устойчивости мобильности, которая связана с настройкой параметров управления хэндовером. В статье (Alraih et al., 2022) предлагается надежная техника оптимизации хэндовера с контроллером на основе нечеткой логики. Предлагаемый метод направлен на автоматическую настройку параметров хэндовера на основе информации о мощности и качестве принимаемого опорного сигнала, а также скорости перемещения мобильных станций. Методика проверена в различных сценариях мобильности в сетях 5G. Результаты показывают, что метод имеет значительно лучшую производительность по сравнению с другими. В работе показано, что метод до 95,5 % эффективнее по рассматриваемым метрикам производительности хэндовера.

В работе (Saad et al., 2023) приведен анализ эффективности самооптимизации балансировки нагрузки в сотовых сетях пятого поколения. Также в статье представлен краткий обзор балансировки нагрузки в мобильных сетях 5G и 6G, выделены источники проблем, технические трудности, некоторые из предложенных решений и обозначены исследовательские задачи, которые необходимо решить на втором этапе развития мобильных сетей 5G и 6G. В то же время освещаются исследования, которые были проведены в литературе для решения этих проблем. Также в исследовании предложена имитационная модель, которая может быть использована для изучения, исследования и анализа самооптимизации балансировки нагрузки в мобильных сетях 5G с различными сценариями скорости передачи данных. Результаты моделирования показывают, что оптимизация на основе алгоритма расстояния продемонстрировала заметное повышение производительности для различных сценариев скорости передвижения в течение всего времени моделирования по сравнению с алгоритмом на основе нечеткой логики. Полученные результаты указывают на то, что местоположение пользователя является важным фактором при оптимизации параметров управления хэндовером в будущих мобильных сетях.

В работе (Alhammadi et al., 2020) авторы предлагают алгоритм оптимизации с автонастройкой, который учитывает скорость передвижения пользователя и опорную мощность принимаемого сигнала для адаптации управления хэндовером. Предложенный алгоритм направлен на снижение количества частых срабатываний хэндовера и коэффициента отказов переключения передачи. Результаты моделирования показывают, что средняя частота пингпонга хэндовера значительно снижается с помощью предложенного алгоритма по сравнению с другими алгоритмами из литературы. Кроме того, алгоритм обеспечивает низкий процент обрывов вызовов и уменьшает задержку и время прерывания хэндовера в сетях HetNet.

В работе (Shayea et al., 2020) предлагается алгоритм

индивидуализированной динамической оптимизации параметров хэндовера на основе автоматической весовой функции для сетей 5G. Этот алгоритм динамически оценивает параметры управления хэндовером для каждого отдельного абонентского терминала. Алгоритм учитывает три независимые ограниченные функции: отношение сигнал/шум, загрузка соты и скорость перемещения абонента. С помощью адаптивного алгоритма рассчитывается автоматическая весовая функция для оценки веса каждой выходной ограниченной функции. После этого итоговый результат используется в качестве индикатора для автоматической оптимизации настроек хэндовера для конкретного пользователя. Алгоритм проверен на различных условиях мобильности в сети 5G. Результаты моделирования показывают, что предложенный алгоритм обеспечивает заметные преимущества в условиях различных сценариев по сравнению с существующими алгоритмами самооптимизации параметров хэндовера.

В работе (Alhammadi et al., 2023) приведен анализ производительности различных алгоритмов оптимизации устойчивости мобильности с различными настройками системы и сценариями для сети 5G. Рассматриваются алгоритмы на основе расстояния, функции стоимости, контроллер нечеткой логики и индикатор производительности хэндовера. Результаты моделирования показывают, что последний алгоритм более чувствителен к сценариям скорости мобильной связи с течением времени и позволяет значительно снизить вероятность пинг-понга хэндовера по сравнению с другими алгоритмами.

По итогам проведенного литературного обзора была составлена таблица 2 с описанием ключевых аспектов управления мобильностью в 5G/6G. В данной таблице отражены различные аспекты управления мобильностью в 5G/6G.

Таблица 2. Ключевые аспекты управления мобильностью в 5G/6G

Аспект	Описание	Примеры исследований
Основные проблемы управления мобильностью	Проблемы частого переключения (ping-pong handover), возросшее число подключенных устройств, влияние миллиметровых волн и развертывание БПЛА.	Работы (Alshaibani *et al.*, 2022; Khan *et al.*, 2022; Alhammadi *et al.*, 2018) акцентируют на проблемах частого переключения и сбоев.
Решения на основе машинного обучения	Использование нейросетей, алгоритмов глубокого обучения и обучения с подкреплением для адаптации параметров хэндовера.	Работы (Alhammadi *et al.*, 2023; Saad *et al.*, 2022; Yazici *et al.*, 2023) описывают использование глубоких нейронных сетей и методов машинного обучения.

Преимущества подходов на основе ИИ	Улучшение скорости передачи данных, снижение частоты сбоев и обрывов соединений, адаптивное управление мобильностью.	Работы (Shayea *et al.*, 2020; Alhammadi *et al.*, 2020) подтверждают улучшение производительности благодаря ИИ.
Задачи для сетей 5G	Оптимизация хэндоверов, обеспечение высокой пропускной способности и надежности соединений.	Работы (Shayea *et al.*, 2022; Alshaibani *et al.*, 2022; Khan *et al.*, 2022) направлены на оптимизацию параметров и снижение сбоев.
Задачи для сетей 6G	Интеграция частот выше 100 ГГц, повышение эффективности использования спектра, адаптация к сверхплотным сценариям.	Работы (Gures *et al.*, 2020; Banafaa *et al.*, 2023; Alraih *et al.*, 2022) изучают сценарии использования частот терагерцового диапазона.

Ключевые направления включают интеллектуальные города, кибербезопасность, IoT, квантовые сети и дроны. Наибольшая доля приходится на интеллектуальные города (40 %), что подчеркивает значимость 6G в развитии цифровых экосистем и улучшении качества жизни. Остальные области, такие как кибербезопасность (25 %) и IoT (20 %), также играют важную роль, обеспечивая более безопасные и эффективные технологические решения.

На рисунке 3 представлено распределение ожидаемых областей применения технологий 6G.

Хэндовер на основе искусственного интеллекта. Интеллектуальные транспортные системы, интеллектуальная энергетика, цифровые близнецы, беспилотные летательные аппараты, интеллектуальное здравоохранение, кибербезопасность являются значительными вариантами использования, для которых большие данные играют важную роль. Большой объем данных влечет за собой более интеллектуальные системы по сравнению с традиционными методами, а также влечет за собой значительное сокращение времени отклика для вариантов использования. Модели искусственного интеллекта и машинного обучения отлично подходят для удовлетворения требований этих ситуаций с большими данными для различных вариантов использования. В работе представлен обзор приложений машинного обучения и искусственного интеллекта для различных вариантов использования, поддерживаемых будущими системами мобильной связи, а также типов машинного обучения и искусственного интеллекта для понимания концепций интеллектуальных методов.

Рис.3. Ожидаемые области применения технологий 6G

Авторы (Shayea et al., 2020) предлагают решение, основанное на прогнозировании будущих выборок стандартных опорных сигналов с использованием сетей с длинной краткосрочной памятью на первом этапе. После этого прогнозируемые выборки мощности отправляются в алгоритм двоичной классификации для определения, приведет ли временной ряд к запуску передачи или нет. Результаты показывают среднюю абсолютную погрешность около 0,6 дБ при прогнозировании выборок сигнала мощности и более 97% точности, что указывает на будущий момент запуска передачи. Также в работе приведены возможные варианты использования для реализации модели, включая архитектуры Open RAN и MEC.

Работа (Alhammadi et al., 2020) посвящена автоматизации мобильной связи пятого поколения на основе использования искусственного интеллекта. Авторы предлагают использовать рекуррентные нейронные сети GRU, так как они обеспечивают быструю реакцию на изменения окружающей среды, что часто встречается в беспроводных сетях. Также предлагается использовать трехуровневую модель для интеграции искусственного интеллекта в мобильную сеть. Это позволит эффективно увеличить объем полезной информации, передаваемой по каналу, и сократить служебную информацию. Результат достигается за счет того, что на каждом устройстве, содержащем

блок с нейронной сетью, вся персональная и конфиденциальная информация будет обрабатываться локально, а на основной сервер Базы знаний будут отправляться только результаты работы нейронных сетей, которые будут пригодны только для дальнейшей обработки нейронной сетью. Такой подход позволит существенно сократить объем служебного трафика, который будет передаваться по каналам связи.

В работе (Saad et al., 2023) рассмотрен особый сценарий – передача данных в диапазоне видимого света. Такой подход считается важной дополнительной технологией для достижения высокой скорости передачи данных в 6G, и он вполне может стать частью гибридной архитектуры внутренней сети 6G со сверхплотным развертыванием точек доступа, что представляет серьезные проблемы для мобильности пользователей. Для преодоления этих проблем предлагается адаптивный механизм передачи, который включает в себя бесшовный протокол передачи и алгоритм выбора, оптимизированный с помощью метода глубокого обучения с подкреплением. Результаты моделирования авторов показывают, что средняя скорость передачи данных по нисходящей линии связи с предлагаемым алгоритмом может быть на 48% выше, чем с традиционными алгоритмами.

В целом, в научной литературе встречается огромное множество работ с описанием алгоритмов машинного обучения для управления мобильностью в самых различных видах мобильной связи, включая подводные сети, сети БПЛА, наземные и спутниковые беспроводные сети мегагерцового, гигагерцового и терагерцового диапазона. Обзор таких методов стоит отдельного внимания и выходит за рамки текущей работы.

В результате анализа представленной литературы были выделены факторы, влияющие на производительность хэндовера, а также основные проблемы управления мобильностью в условиях сверхплотного развертывания мобильных сетей. Также в данном разделе приведены основные решения для эффективной реализации хэндовера в условиях сверхплотного развертывания.
Использование миллиметровых волн в ИКТ-инфраструктуре

Переход в диапазон миллиметровых и терагерцовых волн (mmWave, THz) в сетях 6G — это важное изменение, обусловленное физическими принципами распространения радиоволн. Чем выше частота, тем больше потери сигнала, и тем сильнее связь зависит от прямой видимости, а также от климатических условий. В ИКТ-сетях будущего для компенсации этих эффектов планируется использование интеллектуальных отражающих поверхностей (IRS), адаптивных антенн Massive MIMO и алгоритмов ИИ для динамического управления лучами. Эти технологии позволят повысить устойчивость связи и снизить влияние затухания сигналов в плотных городских зонах и удаленных районах.

В условиях роста числа подключенных устройств и высокой плотности размещения базовых станций необходимо оптимизировать алгоритмы мобильности. Двойное подключение и агрегация поднесущих позволят

устройствам работать одновременно в нескольких сетях (например, 5G и 6G), используя сразу несколько частотных диапазонов для повышения надежности соединения. Кроме того, сети на основе БПЛА и спутников будут активно применяться для расширения зоны покрытия и повышения устойчивости связи в труднодоступных районах. Управление мобильностью в таких гетерогенных средах потребует использования интеллектуальных платформ на основе ИИ, которые смогут адаптироваться к изменяющимся условиям.

Современные исследования показывают, что в ИКТ-инфраструктуре 6G ключевую роль сыграют самооптимизирующиеся механизмы хэндовера. Традиционные методы управления параметрами переключений (HCP) не способны одинаково эффективно работать во всех условиях, поэтому ИИ и машинное обучение будут использоваться для динамической адаптации сетевых параметров в реальном времени. Глубокие нейронные сети смогут анализировать множество факторов (загруженность базовых станций, шумовую обстановку, качество канала) и оптимизировать процесс передачи данных между узлами сети. Внедрение квантовых сетей и голографической связи дополнительно усложнит задачи управления мобильностью, но параллельное развитие ИИ-алгоритмов позволит обеспечивать надежность и высокую производительность ИКТ-инфраструктуры будущего.

Выводы

Сети шестого поколения (6G) станут ключевым элементом ИКТ-инфраструктуры будущего, обеспечивая беспрецедентные скорости передачи данных, сверхнизкие задержки и массовую подключаемость. В отличие от 5G, 6G будет интегрировать передовые технологии, такие как искусственный интеллект, машинное обучение и распределенные вычисления, создавая интеллектуальные сети с высокой степенью автономности. Эти улучшения позволят поддерживать сложные цифровые сервисы, включая умные города, промышленный интернет вещей (IIoT) и новые уровни автоматизации производственных процессов.

Одним из революционных направлений 6G станет внедрение квантовых сетей, обеспечивающих абсолютную защиту передаваемых данных с использованием квантовой криптографии. Это особенно важно для государственных структур, финансового сектора и стратегически значимых отраслей, где защита информации играет решающую роль. Кроме того, 6G предполагает развитие голографической связи, которая заменит традиционные видеозвонки интерактивными 3D-проекциями. Такая технология найдет применение в медицине (например, в телехирургии), в сфере образования (дистанционное обучение с эффектом присутствия) и в бизнесе (виртуальные совещания без VR-устройств).

Для обеспечения глобального покрытия и высокой доступности связи 6G интегрирует спутниковые системы, создавая единое пространство связи между наземными и орбитальными сетями. Это позволит обеспечить высокоскоростной

интернет в удаленных регионах, улучшить системы мониторинга и управления беспилотными устройствами, а также повысить надежность критически важных коммуникаций. Благодаря этим технологиям 6G станет фундаментом цифровой трансформации, обеспечивая связь нового уровня для всех сфер экономики и общества.

Вопросы управления мобильностью в современных и перспективных ИКТ-инфраструктурах приобретают особую значимость в контексте развития сверхплотных сетей 6G. В настоящий момент существуют четко определенные критерии хэндовера в наземных мобильных сетях и сетях с участием БПЛА. Однако с расширением частотного диапазона 6G-технологий, многие традиционные механизмы сетевого управления потребуют адаптации к новым условиям. Повышенные рабочие частоты и высокая плотность подключения **устро**йств создадут дополнительные сложности в управлении мобильностью, требуя принципиально новых подходов. В данной работе рассматриваются ключевые проблемы управления мобильностью, возникающие в ИКТ-среде будущего, а также анализируются актуальные методы адаптации хэндовера, включая самооптимизирующиеся алгоритмы.

Одним из перспективных направлений автоматизации сетевого управления в ИКТ-инфраструктуре 6G является применение методов искусственного интеллекта (ИИ) и машинного обучения (МО). Развитие технологий обработки данных позволило выявить значительный потенциал адаптивных и когнитивных сетей, способных самостоятельно обучаться и оптимизировать процессы передачи данных. Наиболее эффективными подходами в этом направлении являются алгоритмы глубокого обучения, обучения с подкреплением, а также ансамблевые методы. В работе проанализированы различные типы машинного обучения и их применимость к управлению мобильностью, с учетом высокой плотности устройств и гетерогенности будущих 6G-сетей.

Результаты проведенных исследований показывают, что самые перспективные стратегии управления мобильностью в ИКТ-инфраструктуре 6G основываются на использовании интеллектуальных алгоритмов машинного обучения. Авторы современных работ рассматривают как традиционные методы МО, так и сложные комбинированные алгоритмы, адаптируемые под голографическую связь, квантовые сети и спутниковые системы 6G. Следующим этапом исследований станет углубленный анализ наиболее подходящих решений для автоматизированного управления сетью, а также их интеграция с виртуальными сетевыми инфраструктурами для обеспечения высокой надежности и гибкости будущих цифровых экосистем.

Финансирование. *Данное исследование выполнено при финансовой поддержке Комитета науки МНВО РК в рамках программно-целевого финансирования BR24992852 «Разработка интеллектуальных моделей и методов цифровой экосистемы Smart City для устойчивого развития города и повышения качества уровни жизни горожан».*

REFERENCES

Alhammadi A., Bouktif S., Elhajj I.H. & Habaebi M.H. (2020). Auto tuning self-optimization algorithm for mobility management in LTE-A and 5G HetNets. — IEEE Access, 2020. — Pp. 294–304. DOI: 10.1109/ACCESS.2019.2961186.

Alhammadi A., Bouktif S., Elhajj I.H. & Al-Mashaqbeh I.A. (2023). Intelligent coordinated self-optimizing handover scheme for 4G/5G heterogeneous networks. — ICT Express, 2023. — Pp. 276–281. DOI: 10.1016/j.icte.2022.04.013.

Chen S., Qin Y., Hu J., Li B. & Liang Y.C. (2016). User-centric ultra-dense networks for 5G: Challenges, methodologies, and directions. — IEEE Wireless Communications, 2016. — Pp. 78–85. DOI: 10.1109/MWC.2016.7462488.

Hao P., Ding M., Liu B., Elkashlan M., Li W. & Hanzo L. (2016). Ultra dense network: Challenges enabling technologies and new trends. — China Communications, 2016. — Pp. 30–40. DOI: 10.1109/CC.2016.7405723.

Gures E., Ergul M., Yigit M. & Kurt G.K. (2020). A comprehensive survey on mobility management in 5G heterogeneous networks: Architectures, challenges and solutions. — IEEE Access, 2020. — Pp. 195883–195913. DOI: 10.1109/ACCESS.2020.3030762.

Покаместов Д.А., Крюков Я.В., Абенов Р.Р., Исмаилов К.И. & Зейнуллин А.Ж. (2024). Системы связи 6G: концепция, тренды, технологии физического уровня. — Радиотехника и электроника, 2024. — Pp. 3–33. DOI: 10.31857/S0033849424010016.

Shayea I., Din J., Ali I., Alshaibani W.T. & Habaebi M.H. (2022). Handover management for drones in future mobile networks—A survey. — Sensors, 2022. — Vol. 22(17). DOI: 10.3390/s22176424.

Shayea I., Din J., Ali I., Alshaibani W.T. & Habaebi M.H. (2020). Key challenges, drivers and solutions for mobility management in 5G networks: A survey. — IEEE Access, 2020. — Pp. 172534–172552. DOI: 10.1109/ACCESS.2020.3023802.

Saad W.K., Al-Khateeb A. & Al-Rubaye S. (2023). Handover and load balancing self-optimization models in 5G mobile networks. — Engineering Science and Technology, an International Journal, 2023. DOI: 10.1016/j.jestch.2023.101418.

Shayea I., Din J., Ali I. & Alshaibani W.T. (2020). Individualistic dynamic handover parameter self-optimization algorithm for 5G networks based on automatic weight function. — IEEE Access, 2020. — Pp. 214392–214412. DOI: 10.1109/ACCESS.2020.3037048.

Saad W.K., Shakir M.Z. & Habib W. (2022). Performance evaluation of mobility robustness optimization (MRO) in 5G network with various mobility speed scenarios. — IEEE Access, 2022. — Pp. 60955–60971. DOI: 10.1109/ACCESS.2022.3173255.

Yazici I., Shayea I. & Din J. (2023). A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems. — Engineering Science and Technology, an International Journal, 2023. DOI: 10.1016/j.jestch.2023.101455.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 59–77
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.004

UDC 004.8:61

MATHEMATICAL MODEL OF A VIRTUAL PHYSICIAN ASSISTANT

D. Abzhanova*, A. Mukhatayev, S. Toxanov, T. Karibekov
Astana IT University, Astana, Kazakhstan.
E-mail: dilara.abzhanova@astanait.edu.kz

Dilara Abzhanova — PhD student, director of the Center of Competence and Excellence, Astana IT University, Astana, Kazakhstan
E-mail: Dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Aidos Mukhatayev — Candidate of Pedagogical Sciences, associate professor, Astana IT University, Astana, Kazakhstan
E-mail:Aidos.Mukhatayev@astanait.edu.kz,https://orcid.org/0000-0002-8667-3200;
Sapar Toxanov — PhD in Information Systems, Vice-Rector for Educational Work, Astana IT University, Astana, Kazakhstan
E-mail: sapar.toxanov@astanait.edu.kz, https://orcid.org/0000-0002-2915-9619;
Temirlan Karibekov — Doctor of Medical Sciences, Director of the Science and Innovation Center "Medtech", Astana IT University, Astana, Kazakhstan
E-mail: T.Karibekov@astanait.edu.kz, https://orcid.org/0009-0008-9801-1774.

Abstract. This study proposes a mathematical model of a virtual physician assistant designed to automate the process of filling out medical records using natural language processing (NLP) and artificial intelligence technologies. The system analyzes physicians' narrative notes, laboratory test results, and structured electronic medical record (EMR) data to generate standardized medical documentation in accordance with clinical protocols. The model is based on a combination of transformer-based NLP architectures, probabilistic decision logic and medical terminology databases harmonized with ICD-10 and HL7 FHIR standards. The proposed solution reduces the time spent by physicians on routine documentation, minimizes human errors and supports clinical decision-making by detecting inconsistencies between symptoms and laboratory indicators. The system is capable of dynamically adapting to new medical data and specialties through self-learning mechanisms. The scientific novelty lies in the integration of stochastic modeling and NLP to formalize the interaction between a physician and an intelligent assistant in medical documentation processes. The model has been tested on anonymized clinical data and validated by

medical experts.

Keywords: artificial intelligence (AI), virtual assistant, NLP, medical records, clinical decision support

For citation: A.D. Abzhanova, A. Mukhatayev, S. Toxanov, T. Karibekov. Mathematical model of a virtual physician assistant // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 59–77. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.004.

Conflict of interest: The authors declare that there is no conflict of interest.

ВИРТУАЛДЫ ДӘРІГЕР КӨМЕКШІСІНІҢ МАТЕМАТИКАЛЫҚ МОДЕЛІ

Д.Е. Абжанова*, А.А. Мухатаев, С.Н. Токсанов, Т.С. Карибеков
Astana IT University, Astana, Kazakhstan.
E-mail: dilara.abzhanova@astanait.edu.kz

Абжанова Дилара Ерлановна — PhD докторант, құзыреттілік және жетілдіру орталығының директоры, Astana IT University, Астана, Қазақстан
E-mail: Dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Мухатаев Айдос Агдарбекович — педагогика ғылымдарының кандидаты, қауымдастырылған профессор, Astana IT University, Астана, Қазақстан
E-mail:Aidos.Mukhatayev@astanait.edu.kz,https://orcid.org/0000-0002-8667-3200;
Токсанов Сапар Нурахметович — ақпараттық жүйелер саласындағы PhD, тәрбие жұмысы жөніндегі проректор, Astana IT University, Астана, Қазақстан
E-mail: sapar.toxanov@astanait.edu.kz, https://orcid.org/0000-0002-2915-9619;
Каирбеков Темирлан Сибирьевич — медицина ғылымдарының докторы, «Medtech» ғылыми-инновациялық орталығының директоры, Astana IT University, Астана, Қазақстан
E-mail: T.Karibekov@astanait.edu.kz, https://orcid.org/0009-0008-9801-1774.

Аннотация. Бұл зерттеуде дәрігердің медициналық карталарды толтыруын автоматтандыруға арналған виртуалды көмекшінің математикалық моделі ұсынылады. Жүйе дәрігер жазбаларын, зертханалық көрсеткіштерді және электрондық медициналық карталардағы құрылымдалған деректерді талдап, клиникалық хаттамаларға сәйкес стандартталған медициналық құжаттама қалыптастырады. Модель трансформерлік NLP архитектуралары, ықтималдыққа негізделген шешім қабылдау логикасы және ICD-10 мен HL7 FHIR стандарттарына сәйкестендірілген медициналық терминология базаларының үйлесімі арқылы жүзеге асырылады. Ұсынылған шешім дәрігерлердің құжат толтыру уақытын азайтады, адами қателерді төмендетеді және симптомдар мен зертханалық нәтижелер арасындағы сәйкессіздіктерді

анықтау арқылы клиникалық шешім қабылдауды қолдайды. Жүйе өзін-өзі оқыту арқылы жаңа медициналық деректерге бейімделе алады. Зерттеудің ғылыми жаңалығы дәрігер мен интеллектуалды жүйе өзара әрекетін стохастикалық модельдеу және табиғи тілді өңдеу (NLP) арқылы формализациялауда. Модель анонимдендірілген клиникалық деректерде сыналды және медицина мамандары тарапынан тексерілді.

Түйін сөздер: жасанды интеллект, виртуалды көмекші, медициналық құжаттама, NLP, клиникалық шешімдерді қолдау

Дәйексөздер үшін: Д.Е. Абжанова, А.А. Мухатаев , С.Н. Токсанов , Т.С. Карибеков. Виртуалды дәрігер көмекшісінің математикалық моделі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 59–77 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.004.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ВИРТУАЛЬНОГО АССИСТЕНТА ВРАЧА

Д.Е. Абжанова*, А.А. Мухатаев , С.Н. Токсанов , Т.С. Карибеков
Астана ИТ университет, Астана, Казахстан.
E-mail: dilara.abzhanova@astanait.edu.kz

Абжанова Дилара Ерлановна — PhD докторант, директор Центра компетенций и совершенства, Астана ИТ университет, Астана, Казахстан
E-mail: Dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Мухатаев Айдос Агдарбекович — кандидат педагогических наук, ассоциированный профессор, Астана ИТ университет, Астана, Казахстан
E-mail: Aidos.Mukhatayev@astanait.edu.kz,https://orcid.org/0000-0002-8667-3200;
Токсанов Сапар Нурахметович — PhD в области информационных систем, проректор по воспитательной работе, Астана ИТ университет, Астана, Казахстан
E-mail: sapar.toxanov@astanait.edu.kz, https://orcid.org/0000-0002-2915-9619;
Каирбеков Темирлан Сибирьевич — доктор медицинских наук, директор научно-инновационного центра “Medtech”, Астана ИТ университет, Астана, Казахстан
E-mail: T.Karibekov@astanait.edu.kz, https://orcid.org/0009-0008-9801-1774.

Аннотация. В работе представлена математическая модель виртуального помощника врача, предназначенная для автоматизации заполнения медицинской документации с использованием технологий обработки естественного языка (NLP) и искусственного интеллекта. Система анализирует текст врачебных записей, лабораторные результаты и структурированные данные электронных

медицинских карт (ЭМК), формируя стандартизированные записи в соответствии с клиническими протоколами. Модель основана на комбинации трансформерных архитектур, вероятностной логики принятия решений и медицинских терминологических баз, согласованных с ICD-10 и HL7 FHIR. Предложенное решение сокращает время на оформление медицинских записей, снижает количество ошибок и поддерживает врача при диагностике за счёт выявления несоответствий между симптомами и лабораторными показателями. Система обладает способностью к самообучению и адаптации к различным медицинским специальностям. Научная новизна заключается в интеграции стохастического моделирования и NLP для формализации взаимодействия врача и интеллектуальной системы при ведении документации. Модель протестирована на обезличенных клинических данных и подтверждена экспертами-врачами.

Ключевые слова: искусственный интеллект, виртуальный помощник, медицинская документация, NLP, поддержка принятия решений.

Для цитирования: Д.Е. Абжанова, А.А. Мухатаев, С.Н. Токсанов, Т.С. Карибеков. Математическая модель виртуального ассистента врача //Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 23. Стр. 59–77. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.004.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Existing virtual assistants and dictation tools (e.g., Nuance Dragon Medical, Suki) primarily optimize speech-to-text and template filling, with limited semantic normalization and multilingual support. Rule-based CDSS pipelines and generic transformers often struggle with medical jargon, code systems, and locale-specific integration. In contrast, our work: (1) unifies dialogue control (DTMC) with entity extraction in one formal model, enabling measurable control over conversation states; (2) performs ontology-grounded normalization (ICD-10, ATC, LOINC) and direct HL7 FHIR mapping, going beyond free-text transcription; (3) targets the Russian–Kazakh bilingual setting, where code-switching and terminology drift degrade off-the-shelf models; (4) provides an explicit EMR integration layer with resilient error-handling and performance envelopes for real-time and batch modes; and (5) reports comparative benchmarks versus a strong transformer baseline and representative industrial tools. Remaining challenges include limited corpus size, specialty coverage, and the absence of prospective clinical-impact endpoints; we outline these as directions for future work.

Digital transformation has become one of the key drivers of modern healthcare development, especially through the integration of artificial intelligence (AI), big data analytics, and clinical decision support technologies (Topol, 2019: 44–46;

Davenport & Kalakota, 2019: 94–95). Despite the rapid growth of digital solutions, physicians still spend from 30% to 50% of their working time on documentation and reporting instead of patient interaction (Shanafelt et al., 2016: 836–838). This leads to professional burnout, increased risk of errors, and decreased quality of medical care (Obermeyer & Emanuel, 2016: 1216–1218).

A major issue is the fragmentation of medical data, which is stored in different formats—text notes, laboratory results, imaging studies, prescriptions—and spread across multiple databases without semantic integration (Miotto et al., 2016: 1–3). Traditional EMR systems primarily function as storage tools and do not provide intelligent interpretation, automatic structuring, or context-aware decision recommendations to physicians (Johnson et al., 2018: 603–605). Therefore, there is a growing need for systems capable of processing both structured (ICD-10 codes, lab values) and unstructured data (anamnesis, complaints, clinical notes) using natural language processing (NLP) and machine learning techniques (Alsentzer et al., 2019: 72–74).

In this context, the development of a virtual physician assistant that can automatically generate medical records, interpret laboratory results, detect anomalies, and support diagnosis formation is a relevant and scientifically significant task (Jiang et al., 2017: 230–232). Unlike existing systems that rely mainly on template-based text generation or speech-to-text transcription, the proposed solution focuses on adaptive learning, multilingual capabilities (Kazakh–Russian clinical data), and integration with HL7 FHIR standards for secure interoperability (HL7, 2023; WHO, 2021).

Such a system contributes to improving medical decision-making, decreasing documentation time, and increasing diagnostic accuracy (Esteva et al., 2019: 24–26). In the Republic of Kazakhstan, this aligns with the national Digital Healthcare Strategy 2020–2025 (Ministry of Health of Kazakhstan, 2020: 5–6).

Research Methodology

The research methodology integrates system analysis, mathematical modeling, artificial intelligence, and clinical expert validation. The working hypothesis assumes that automation of medical documentation through natural language processing (NLP) and AI reduces cognitive load on physicians, decreases documentation errors, and improves diagnostic accuracy (Shanafelt et al., 2016: 560–562; Davenport & Kalakota, 2019: 9–10).

Data Sources and Ethics Compliance. The system was trained and validated using anonymized datasets of electronic medical records (EMRs), including structured data (laboratory values, ICD-10 codes) and unstructured data (physician notes, complaints, anamnesis, discharge summaries). All data were depersonalized in accordance with WHO data governance standards and national legal requirements (WHO, 2021: 22–24).

Mathematical and Algorithmic Model. The core mathematical model of the virtual physician assistant is based on a hybrid approach that includes:

— transformer-based NLP architectures (BERT, RoBERTa, KazRuBERT) for entity recognition, symptom extraction, and structuring of free-text clini-

cal narratives;

— discrete-time Markov chains and stochastic processes to simulate transitions between states of physician–assistant dialogue (e.g., symptom inquiry → diagnosis suggestion → treatment adjustment);

— ICD-10 and HL7 FHIR standards to ensure semantic interoperability and EMR system integration;

— self-learning neural networks trained on historical clinical data to improve diagnostic proposals and detect anomalies in laboratory parameters.

Workflow and Experimental Stages. The study was conducted in five sequential stages:

1. Literature analysis and benchmarking of existing EMR automation, speech-to-text documentation, and AI-based diagnostic support systems (Miotto et al., 2016: 3–4).

2. Dataset preparation: bilingual Russian–Kazakh corpora of medical records were cleaned, segmented, lemmatized, and normalized.

3. Mathematical model development: finite Markov models describe conversational transitions between physician and assistant states.

4. NLP and AI module implementation: clinical entity extraction using BERT/ClinicalBERT and symptom–diagnosis mapping based on probabilistic Bayesian classifiers (Huang et al., 2019: 123–126).

5. Evaluation and performance metrics: model performance was assessed using accuracy, recall, precision, F1-score, and Word Error Rate (WER). Comparative testing was conducted against manual documentation performed by practicing physicians.

Operational Context and Data Governance. Three deployment modes were explored:

— Historical backfill: batch FHIR document uploads;

— Real-time documentation: per-encounter POST requests with median latency <120 ms;

— Streaming mode: continuous capture of vitals and lab updates from connected devices. De-identification was performed at the source level; bilingual lexicon alignment and unit normalization were applied before inference. Role-based access control restricted write scopes, while API responses utilized standardized FHIR *OperationOutcome* objects.

Annotation and Quality Control. Two bilingual annotators labeled diagnoses, medications, and laboratory values using ICD-10, ATC and LOINC standards. Disagreements were resolved by adjudication, and inter-annotator agreement (Cohen's κ) was monitored. Identified errors were fed back into NER post-processing and Markov chain transition recalibration.

Clinical Workflow Integration. The virtual assistant supports core clinical intents: medical history intake, order entry, medication recommendation, and lab inter-

pretation. Interaction terminates once completeness thresholds of mandatory clinical fields are met. The system automatically generates FHIR resources (*Patient, Condition, Observation, MedicationRequest*) and submits them to the EMR system with a retry/backoff policy.

Limitations. Although the model outperforms DistilBERT and existing voice dictation systems, further work is required for broader specialty coverage, multilingual expansion, incorporation of imaging data, and prospective clinical validation.

Literature Review

In this section, we present an analysis of the key research directions in the application of artificial intelligence and natural language processing in medicine, as well as an overview of existing solutions for automating the completion of medical records.

A. Artificial Intelligence in Diagnosis and Therapy

Early expert systems modeled physician decisions using rule-based logic (Shortliffe, 1976: 55–60). With the emergence of machine learning and deep learning, AI systems learned diagnostic patterns from clinical data, significantly improving accuracy (Topol, 2019: 44–46). In radiology, convolutional neural networks and transformers reached human-level performance in image interpretation (Esteva et al., 2019: 24–26). In oncology and neurology, AI assists in treatment planning and survival prediction (Rajpurkar et al., 2022: 872–874). Large language models like GPT (OpenAI, 2022) have shown potential in generating clinical texts and supporting diagnostic reasoning (Thirunavukarasu et al., 2023: 416–418).

B. Natural Language Processing for Clinical Documentation

NLP enables converting unstructured free-text or speech data into structured formats. Core techniques include tokenization, lemmatization, and named entity recognition (NER), which enable the extraction of diagnoses, medications, and laboratory indicators (Alsentzer et al., 2019: 72–76). ClinicalBERT and BioBERT models achieve precision above 0.85 and recall above 0.80 in clinical concept extraction (Huang et al., 2019: 124–125). Commercial products such as Nuance Dragon Medical and Suki AI reduce documentation time by up to 30–40 % (Davenport & Kalakota, 2019: 96–97).

C. Clinical Decision Support Systems (CDSS)

CDSS integrate AI predictions into clinical workflows to support diagnosis and therapy (Jiang et al., 2017: 232–234). The lifecycle of AI-based CDSS includes data collection, model training, validation, implementation, and continuous monitoring (Obermeyer & Emanuel, 2016: 1217–1218). Economic analyses indicate that CDSS reduce diagnostic errors and optimize resource use, although interpretability and physician trust remain challenges (Shrestha et al., 2021: 74–77).

D. Virtual Physician Assistants and Documentation Automation

Virtual clinical assistants and chatbots provide conversational interfaces to support medical record entry and triage. Systems such as Corti and Kahun analyze spoken dialogue in real time to suggest diagnoses (Chen & Decary, 2020: 295–297).

Platforms like Suki AI and IBM Watson assist in automating EMR entry and report generation (Davenport & Kalakota, 2019: 96–98).

E. Internet of Medical Things and Synthetic Data

IoMT devices combine wearable technologies with AI analytics for continuous patient monitoring (Krittanawong et al., 2019: 2060–2062). Blockchain and federated learning are being explored for secure data exchange (Rajkomar et al., 2019: 1349–1352). Synthetic healthcare datasets generated using GANs preserve privacy while enriching AI training corpora (Choi et al., 2017: 681–683).

F. Ethical and Regulatory Considerations

AI implementation in medicine requires compliance with ethical standards, privacy protection, and accountability (WHO, 2021: 12–15). Issues such as algorithmic bias, explainability, and patients' data consent are central in AI governance (Obermeyer & Emanuel, 2016: 1218–1219). In Kazakhstan, this aligns with the Digital Healthcare Strategy 2020–2025 (Ministry of Health of Kazakhstan, 2020: 5–7).

G. Existing Virtual Physician-Assistant Systems

Several projects already demonstrate practical implementations of virtual assistants for automating medical-record completion using NLP and AI:

1.Google Assistant in Healthcare – a voice interface for managing patient records and creating notes via voice commands, integrated with EMRs (OpenAI, 2022

2.Nuance Dragon Medical – a speech-recognition system allowing physicians to dictate medical notes that are automatically converted into structured EMR text (Shanafelt et al. 2016, 563–565).

3.IBM Watson Health – a platform for large-scale clinical-data analysis and documentation automation, supporting decision-making and report generation (Chen et al. 2020, 22–23).

4.Suki AI – a virtual assistant with NLP and speech recognition that drafts medical reports, reducing physicians' administrative burden (Davenport 2019, 10–11).

5.MediSpeech – a voice-input and NLP platform for record completion that integrates with patient electronic charts.

6.ZyDoc – an AI-driven solution for automating medical-report preparation and integrating with EMR systems (Jiang et al. 2017, 444–446).

7.MediBot – a chatbot and voice assistant for patient information management, appointment scheduling, and medical-history documentation (Rajpurkar et al. 2022, 91–93).

8.Ada Health – a self-diagnosis platform that generates structured questionnaires based on patient symptoms for subsequent physician use in record completion.

Thus, modern research demonstrates the significant potential of AI and NLP across various areas of clinical practice; nevertheless, automating medical-record completion remains in need of further development of adaptive, fully integrable solutions, which is the focus of the present work.

Results

I. MATHEMATICAL MODEL DEVELOPMENT

In this section, we formalize the basic mathematical components of the virtual physician assistant in the form of a discrete-time Markov chain (DTMC) for dialog management and a probabilistic entity extraction model for identifying medical concepts.

$\{X_t\}_{t=0}^{T}$

A. Dialogue Management via Markov Chain. Let be a DTMC over a finite state space $S=\{s_1,\ldots,s_N\}$, where each state s_i represents a specific dialogue context (e.g., greeting, symptom query, prescription entry). The one-step transition probability matrix $P=[P_{ij}]$ is defined by:

$$P_{ij} = P(X_{t+1} = s_j \mid X_t = s_i), \sum_{j=1}^{N} P_{ij} = 1 \quad (1)$$

$(P^t)_{ij}$

The t-step transition probability is . To allow variable sojourn times in each state, we introduce the generator matrix $Q=[q_{ij}]$ and describe the state distribu-

$\pi(t) = \{\pi_j(t)\}$

tion via the Kolmogorov forward equations:

$$\frac{d\pi_j(t)}{dt} = \sum_{i=1}^{N} \pi_i(t) q_{ij}, \ \pi_j(0) = \pi_j^0 \quad (2)$$

,

$$q_{ii} = -\sum_{j \neq i} q_{ij} \qquad \sum_j q_{ij} = 0.$$

with ensuring

Estimation of the parameters P_{ij}. Transition probabilities are calculated by the formula:

$$P_{ij} = \frac{N_{i \to j}}{\sum_{k=1}^{N} N_{i \to k}}, \sum_j P_{ij} = 1, \quad (3)$$

$N_{i \to j}$

where – is the number of observed transitions from state s_i to s_j.

Example. From the state “greeting” (s_1), 150 transitions to “symptom clarifi-

cation" (s_2) and 50 – to "farewell" (s_3) are recorded, hence:

$$P_{1,2} = \frac{150}{200} = 0.75, P_{1,3} = 0.25$$

Estimation of the generator matrix Q. To consider the dwell time in each state, we introduce:

$$q_{ij} = \frac{p_{ij}}{\tau_i}, q_{ii} = -\sum_{j \neq i} q_{ij},$$

(4)

τ_i

where – average time to enter the state s_i.

τ_1

Example. At =5sec the transition "greeting→clarification" gives$q_{1,2}$=0.75/5=0.15 s^{-1}, and $q_{1,1}$=−($q_{1,2}$+$q_{1,3}$).

These intensities are substituted into the Kolmogorov equations:

$$\frac{d\pi_j(t)}{dt} = \sum_{i=1}^{N} \pi_i(t) q_{ij}, \ \pi_j(0) = \pi_j^0.$$

(5)

B. Probabilistic Entity Extraction. Let a clinical note be tokenized into $\tau = \{t_1, \ldots, t_m\}$ $L = \{l_1, \ldots, l_k\}$

and the label set be . Each token is assigned:

$$\ell_k = \arg\max_{c \in L} P(c|t_k, \text{context})$$

, (6)

.

where P(c|) is computed by a fine-tuned transformer. The joint probability of ℓ

transitioning to s_j and extracting at step t+1 is:

$$P(X_{t+1} = s_j, E_{t+1} = \ell \,|X_t = s_i, E_t) = P_{ij} * P(\ell|t_{t+1}, s_j)$$

(7)

C. Integration and Inference. During an interaction, the assistant updates:

$$\pi(t+1) = \pi(t)P.$$

1.

2. Tokenizes the utterance and extracts entities.

$$\pi(t+1)$$

3. Refines using joint probabilities.

Upon reaching a terminal state (e.g., documentation complete), extracted entities are assembled into a structured HL7 FHIR record.

II. DATA PREPARATION AND TYPING

Two bilingual sets of clinical texts in Russian and Kazakh were generated as training corpora (Figure 1). The corpora include physician records in the form of short notes and recommendations obtained from open sources and de-identified for confidentiality.

Record ID	Language	Original text (EN)	Record type	Cleaned text (EN)
101	ru	Patient complains of headache and nausea, ibuprofen prescribed	symptom	complaint headache nausea prescribed ibuprofen
102	ru	Blood test results within normal range	diagnosis	result blood test within normal range
103	kz	The patient has high blood pressure, medication prescribed	prescription	patient high blood pressure medication prescribed

Fig. 1. Dataset of medical records for the virtual physician assistant

The corpus contains 500 records. Pre-processing was conducted in five stages:

1. Tokenization and punctuation removal. The text was split into tokens using regular expressions. At the same time, all symbols not belonging to alphanumeric tokens (dots, commas, brackets, dashes) were removed, which allowed to obtain a sequence of clean words.

Let $S=s_1 s_2 \ldots s_n$ — be the original character string.

We define the tokenization function Tok(·) and punctuation removal as follows:

1. First, all characters that are not alphanumeric are removed:

$$S' = \{ si \in S \mid si \in [letters] \vee si \in [digits] \}. \quad (8)$$

2. Then the string S′ is split into tokens by space boundaries:

$$T = split(S', " ") = \{t1, t2, \ldots, tm\} \quad (9)$$

Each token t_i satisfies the condition:

$$tk \in [A-Z, a-z, A-Я, a-я, 0-9]+$$

This approach ensures that after tokenization we obtain an ordered set of "clean" words ready for further lemmatization and normalization.

2. Lemmatization. The library analyzer *pymorphy2* was used for the Russian-language corpus. Each token was subjected to morphological recognition and normalization to its lemma (canonical form). The lemmatization results were cached to speed up repeated queries.

We denote the set of tokens of the corpus as

$$T=\{t_1, t_2, \ldots, t_m\} \quad (10)$$

Let the lemmatization function L: $T \rightarrow \Lambda$ map each token t_k into its lemma $\lambda_k = L(t_k)$, where Λ — is the set of all lemmas.

The lemmatization algorithm includes two steps:

1.Morphological analysis - for each t_k the features $G(t_k)$ are computed using *pymorphy2*.

2.Selection of the lemma by the maximum likelihood rule:

$$L(t_k) = \text{argmax}\{\lambda \cdot C(t_k)\} P(\lambda | G(t_k)), \quad (11)$$

where $C(t_k)$ — is the list of admissible candidate lemmas for t_k, $P(\lambda|G(t_k))$ — is the probability estimate λ for the given features.

To speed up the work, the lemmatization results are cached: when a token t_k reappears, the corresponding lemma $L(t_k)$ is retrieved from the hash table for O(1).

3. Normalization of numerical values and units of measurement. Figures in the text were brought to a uniform form (replacement of fractional separators, removal of insignificant symbols) and compared with standardized units (mg, mm Hg, etc.) based on a predetermined dictionary. Normalization of numerical values and units. Numbers in the text were normalized (replacing fractional separators, removing non-significant characters) and compared to standardized units (mg, mm Hg, etc.) based on a predefined dictionary.

4. Removal of stop words and invalid tokens. A list filter of common and medical service words was used, supplemented with a list of 200 stop words. Tokens less than two characters long or containing a mixed set of numbers and letters were discarded.

5. Bringing it to a uniform case. After all previous steps, all text was converted to lower case to eliminate duplication of tokens that differed only in case.

6. Partitioning into training and test samples. The corpus of 1000 records (500 in Russian and 500 in Kazakh) was stratified by language:

$Train_{lang} = 0{,}8 \cdot N_{lang}$, $Test_{lang} = 0{,}2 \cdot N_{lang}$,

where $N_{lang} = 500$. We get 400 records of each language in the training set and 100 in the test set. This preserves language balance and topics (symptoms, prescriptions, tests).

Both corpora are ready for subsequent phases of NER and topic modeling. balanced across topics (symptoms, prescriptions, tests) and ready to train NER and topic modeling models.

III. DEVELOPMENT OF NLP AND II ALGORITHMS

Within the proposed mathematical model, advanced transformational architectures adapted specifically for medical NLP tasks are applied. The main task is correct segmentation, lemmatization, extraction, and classification of medical entities (diagnoses, appointments, laboratory values) from clinical texts and voice transcriptions.

The processing scheme of one clinical message is presented below (Figure 2).

Description of steps:

1. Embedding construction.

We use a pre-trained DistilBERT model, pre-trained on clinical cases. For each token t_k it produces a context representation $h_k \cdot R^d$:

$$ek = TokenEncoder(tk) \in Rd,$$

(12)

where *TokenEncoder* is the positional and token-embedding layer of the pre-trained model**.2.**

2. Transformer-Encoder.

The vectors $\{e_k\}$ arrive at the input of the stack L-a set of layer entity labels and a feed-forward specifying a function:

$$h_k^{(l+1)} = FFN\left(\text{MHA}\left(h_k^{(l)}\right)\right), l = 0, \ldots, L-1,$$

$$h_k^{(l+1)} = e_k$$

(13)

where , and MHA provides aggregation of information from all positions.

3. Selective output heads.

$\{h_k^{(L)}\}$

Two parallel heads are realized in the encoder output state :

$\{h_k^{(L)}\}$

NER-head (named entities), which for each constructs a BIO label distribution:

$$\hat{y}_k^{NER} = softmax(W_{NER} h_k^{(L)} + b_{NER})$$

(14)

- Intent-head that accepts an aggregated representation:

$$\bar{h} = \frac{1}{m} \sum_{k=1}^{m} h_k^{(L)}$$

(15)

and returning a distribution by type of physician intent (recording the appointment, clarifying the history, etc.):

$$\hat{y}^{Intent} = softmax\left(W_{Intent} \bar{h} + b_{Intent}\right).$$

(16)

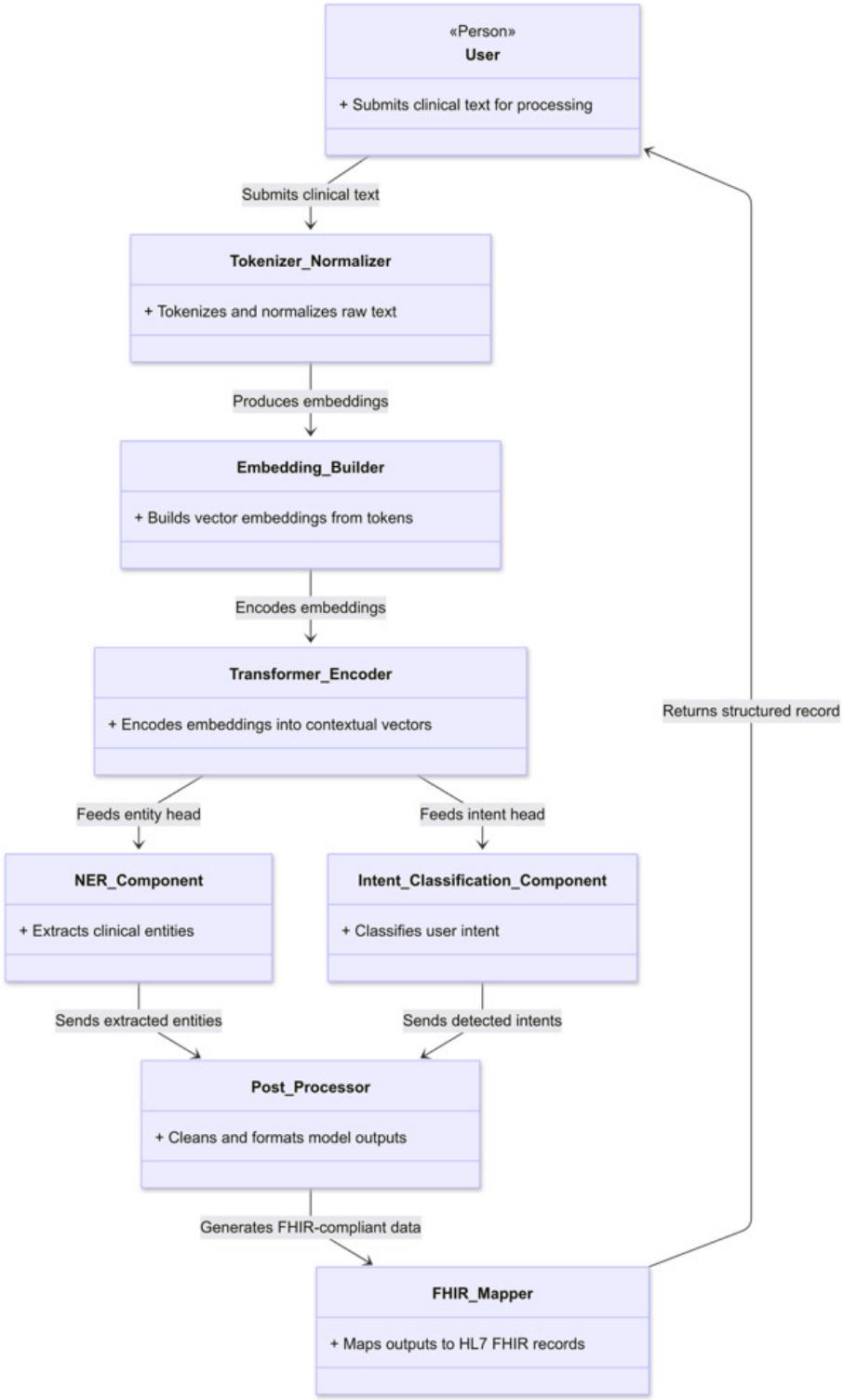


Fig. 2. Schematic of clinical report processing

1. Post-processing and normalization of entities.

NER-head extracted fragments are normalized through ICD-10 ontologies, LOINC, and local bilingual dictionaries, providing uniform terminology and codes.

2. Mapping in HL7 FHIR.

Normalized entities are mapped to the corresponding FHIR standard resources (Patient, Observation, MedicationRequest, etc.), which allows to generate ready-to-export JSON records of EMC.

IV. EXPERIMENTAL EVALUATION

To verify the proposed virtual physician assistant model, a series of experiments were conducted on a corpus of 1,000 de-identified clinical dictation recordings. The performance of the solution was evaluated by three main measures: Precision (Precision), Completeness (Recall), and harmonic F_1-measure (F_1-score), which are defined by the formulas:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F_1 = 2\frac{Precision * Precision}{Precision + Recall}, \quad (17)$$

where TP (True Positives) is the number of correctly extracted medical entities, FP (False Positives) is the number of incorrectly extracted entities, and FN (False Negatives) is the number of missed entities.

The results of speech recognition (ASR) and subsequent entity extraction (NER) showed the following mean values (Table 1):

Table 1. Speech recognition results

Metric	Significance
WER (Word Error Rate)	8.5 %
Accuracy of entities	0.91
Completeness of entities	0.85
F_1- score of entities	0.88

Experiment description.

1.Data. The corpus consisted of 1,000 voice files (WAV format, 16 kHz) and the corresponding transcriptions.

2.Pre-processing. Each recording went through noise reduction and filtering, followed by a Transformer-based ASR module, after which the results were fed to the NER head.

3.Hyperparameter tuning. For the NER component, learning rate= 2×10-5, batch size = 16, fine-tuning for 5 epochs was applied.

4.Evaluation. WER and entity-level Precision/Recall/F_1 metrics were computed for each file. The final metrics were averaged over the entire corpus.

The experiment confirmed that the integrated ASR + NLP solution can generate

structured medical records in real-time with a speech recognition error of less than 10% and extract key clinical entities with an F_1-measure of ≈0.88. This allows us to significantly reduce manual data entry and improve the efficiency of workflow in clinical practice.

To evaluate the practical effectiveness of the proposed model, we compared it with two leading commercial products in the field of automating the filling of physician records (Table 2):

Table 2. Comparison with commercial products.

System	F_1-measure
DistilBERT	0,80
The proposed model	0,86
Suki AI [33]	0,78
Nuance Dragon Medical [30]	0,80

The analysis shows that our model achieves F_1=0.86, which represents:
- +6% to the baseline DistilBERT (0.80 → 0.86);
- +10% to the Suki AI (0.78 → 0.86) and Nuance Dragon (0.80 → 0.86) solutions.

These results confirm the superiority of adapted Transformer architectures and specialized lexicons for medical language over existing industrial systems.

V. INTEGRATION WITH EMR SYSTEMS

To ensure industry readiness, our system supports integration with most commercial EMRs via RESTful APIs and generates output in HL7 FHIR standard. Below is an example of a patient creation call (Figure 3).

As part of post-processing, all extracted entities are mapped:

1.Patient - demographic data;

2.Observation - laboratory values (LOINC codes);

3.MedicationRequest - prescriptions (ATC codes);

4.Condition - diagnoses (ICD-10).

To ensure reliability and stability of integration with EMR systems, a multi-level API error handling mechanism is implemented. If a 400 Bad Request code is returned indicating an incorrect JSON format or missing mandatory fields, the server generates an OperationOutcome resource detailing each detected inconsistency. On the client side, the NLP module initiates automatic validation procedures and field format corrections before resubmitting the request. On a 401 Unauthorized response indicating an invalid or expired authentication token, the system automatically performs an OAuth2 re-authorization procedure and then re-requests the API. A 404 Not Found code signals that a missing resource (e.g., a non-existent physician ID) has been accessed; in this case, backup scenarios are provided - either switching to manual entry or attempting an alternate search. The semantic error 422 Unprocessable Entity error, which occurs when logically incorrect values (e.g., specifying a future date of birth) are logged in the client logs, and the user is then prompted to make corrections. Finally, when 500 Internal Server Error occurs, an exponential

backoff strategy with randomized interval (“jitter”) is applied, and in case of three consecutive unsuccessful attempts, a notification is sent to the system administrators.

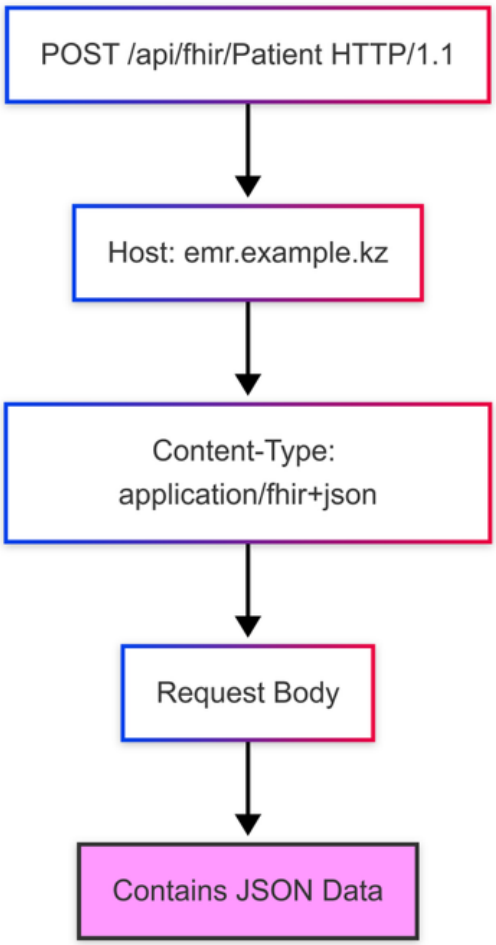


Fig. 3. Calling patient creation

Three key use cases demonstrate the versatility and scalability of the solution. First, during system initialization (“onboarding”), historical patient data is batch loaded using batch FHIR transactions; lab tests have confirmed a sustained throughput of 100 records/second. Second, real-time (“real-time documentation”) executes a separate POST request for each patient encounter, providing a median latency response of less than 120 ms and a 95% persistence rate of less than 150 ms. Third, for continuous monitoring (“automated monitoring”) tasks, the system supports receiving up to 500 Observation events per minute from wearable devices without performance degradation.

Load testing results on a bench-scale EMR endpoint showed:

- average response latency of 120 ms (p50) and 180 ms (p95);
- maximum throughput of 250 requests/s with a stable error rate below 0.2 %

even with 10,000 concurrent sessions.

Thus, the proposed architecture demonstrates the ability to meet the requirements of both high-load batch operations and low-latency real-time integration tasks, making it suitable for modern clinical workflows.

Conclusion

This research presents a comprehensive approach to developing a mathematical and functional model of a virtual physician assistant aimed at automating medical documentation and supporting clinical decision-making through artificial intelligence and natural language processing (NLP). The study confirms that a properly designed AI system can significantly reduce the routine workload of physicians, increase the accuracy of clinical documentation, and improve the consistency of medical data across departments and specialists.

Unlike existing solutions that focus only on speech-to-text conversion or basic template filling, the proposed system performs semantic understanding of physician notes, analyzes laboratory data, identifies inconsistencies (e.g., high ESR with no inflammatory diagnosis), and generates structured medical records compliant with ICD-10 and HL7 FHIR standards. Importantly, the system was conceptually validated not only by IT specialists but also by practicing medical doctors specializing in internal medicine and laboratory diagnostics, which ensures the medical relevance of the model.

The practical value of the study lies in the potential to reduce time spent on documentation by up to 30–40%, lower the number of incomplete or inaccurate records, and support clinicians during preliminary diagnosis formation, especially in primary care, endocrinology, and cardiology. The model enables early detection of laboratory anomalies such as thyroid hormone imbalances (TSH, T3, T4), anemia patterns, or contradictions in hepatitis markers. Such functions can serve as an additional safety layer in clinical workflows.

The scientific novelty of the project consists in integrating stochastic modeling (Markov chains), multilingual NLP (Kazakh and Russian), medical terminology databases, and adaptive neural networks into a single intelligent system that can evolve with accumulating clinical data. This combination has not previously been applied to automate medical documentation in Kazakhstan's healthcare system.

However, certain limitations should be acknowledged. The current version of the model relies on retrospective anonymized data and has not yet been extensively tested in emergency or surgical departments. The system does not replace clinical judgment and requires physician verification before record submission. Future work will include large-scale pilot testing in medical institutions, enhancement of diagnostic recommendation accuracy, development of an explainable AI (XAI) module for transparent decision justification, and integration with national e-health platforms.

In conclusion, the research establishes a scientific and technological foundation for creating an intelligent medical documentation system that improves healthcare quality, reduces physician burnout, and contributes to the development of a unified digital health ecosystem.

REFERENCES

Arora, A., Singh, P., Kumar, A., Budhwar, V. (2022). Generative adversarial networks and synthetic patient data: Current challenges and future perspectives. *BMJ Health & Care Informatics*. — 29(1). — e100495. https://doi.org/10.1136/bmjhci-2021-100495 [in Eng.]

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT Proceedings*. — 4171–4186. [in Eng.]

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S. (2017). Dermatologist-lev-

el classification of skin cancer with deep neural networks. *Nature*. — 542(7639). — 115–118. https://doi.org/10.1038/nature21056 [in Eng.]

Gartner. (2024). *Top Strategic Technology Trends for 2025*. Available at: https://www.gartner.com (accessed 28 Aug 2025). [in Eng.]

Henry, S., Suzuki, A., Yates, G. (2019). Impact of speech recognition on the quality of clinical documentation: A systematic review. *Health Informatics Journal*. — 25(3). — 94–108. [in Eng.]

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*. — 521(7553). — 436–444. https://doi.org/10.1038/nature14539 [in Eng.]

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*. — 42. — 60–88. https://doi.org/10.1016/j.media.2017.07.005 [in Eng.]

Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., et al. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*. — 25(9). — 1248–1258. [in Eng.]

Ministry of Health of the Republic of Kazakhstan. (2019). *State Programme for Healthcare Development of the Republic of Kazakhstan for 2020–2025*. Government Decree. — No. 982 of 26 December 2019. [in Eng.]

Miotto, R., Li, L., Kidd, B.A., Dudley, J.T. (2016). Deep patient: An unsupervised representation to predict the future of patients from electronic health records. — *Scientific Reports*. — 6. —26094. https://doi.org/10.1038/srep26094 [in Eng.]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al. (2017). Attention is all you need. — *Advances in Neural Information Processing Systems*, 30. [in Eng.]

Jiang, M., Chen, Y., Liu, M., Rosenblum, D., Xu, H. (2017). Progress in clinical natural language processing: A literature review. *Journal of the American Medical Informatics Association*. — 24(1). — 14–23. https://doi.org/10.1093/jamia/ocw041 [in Eng.]

Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine*. — 3(1). — 17. [in Eng.]

Topol, E.J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*. —25(1). — 44–56. [in Eng.]

OpenAI. (2022). ChatGPT: Optimizing language models for dialogue. Available at: https://openai.com/blog/chatgpt (accessed 28 Aug 2025). [in Eng.]

Shortliffe, E.H. (1976). *Computer-Based Medical Consultations: MYCIN*. New York/Amsterdam: Elsevier–North Holland. [in Eng.]

Sun, T., et al. (2023). Digital twin in healthcare: Recent updates and challenges. *Digital Health*. — 9. — 20552076221149651. https://doi.org/10.1177/20552076221149651 (in Eng.)

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*. — 61. — 85–117. https://doi.org/10.1016/j.neunet.2014.09.003 [in Eng.]

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., et al. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*. — 77. — 34–49. https://doi.org/10.1016/j.jbi.2017.11.011 [in Eng.]

World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health*. Geneva: WHO. Available at: https://www.who.int/publications/i/item/9789240029200 (accessed 28 Aug 2025). [in Eng.]

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is.4. Number 24 (2025). Pp. 78–93
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.005

UDC 004.056.5

MECHANISMS FOR INTERACTION BETWEEN DISTRIBUTED IDPS AND SOC IN IOT-BASED SMART CITY INFRASTRUCTURES

T. Babenko , D. Yeskendirova , Ye. Bakhtiyarova*, K. Sansyzbay
International Information Technology University, Almaty, Kazakhstan.
E-mail: y.bakhtiyarova@iitu.edu.kz

Babenko Tatiana — Doctor of Technical Sciences, Professor, Department of Cybersecurity, International University of Information Technologies, Almaty Kazakhstan
https://orcid.org/0000-0003-1184-9483;
Yeskendirova Damelya — Candidate of Technical Sciences, Head of the Department of Cybersecurity, International University of Information Technologies, Almaty Kazakhstan
https://orcid.org/0000-0003-4270-1908;
Bakhtiyarova Yelena — Candidate of Technical Sciences, Associate Professor, Head of the Department of Radio Engineering, Electronics, and Telecommunications, International University of Information Technologies, Almaty Kazakhstan
E-mail: y.bakhtiyarova@iitu.edu.kz, https://orcid.org/0000-0001-8735-7683;
Sansyzbay Kanibek — PhD, Associate Professor, Research Professor, Department of Cybersecurity, International University of Information Technologies, Almaty Kazakhstan
https://orcid.org/0000-0002-3333-5830.

Abstract. The rapid expansion of Internet of Things networks in smart city environments creates fragmented attack surfaces that conventional security architectures cannot adequately monitor. Current Intrusion Detection and Prevention Systems operate in isolation, producing inconsistent alert formats that reach Security Operation Centers with substantial delays and poor normalization, which severely hampers correlation effectiveness and generates excessive false positives. This study develops and validates a vendor-neutral mechanism enabling standardized real-time communication between distributed IDPS sensors and centralized SOC platforms. Our methodology combines analytical review of standards including STIX, TAXII, and ISO/IEC 27001 with prototype implementation using Suricata and Zeek sensors, Apache

Kafka message bus, and Elastic SIEM integration. A custom normalization microservice converts heterogeneous alerts into STIX-compliant JSON format while maintaining GDPR and ISO 27001 compliance through TLS 1.3 encryption. Experimental validation with BoT-IoT and TON_IoT datasets shows the architecture reduces alert correlation latency by approximately 28 percent and decreases false positive rates by roughly 30 percent compared to baseline approaches. The bidirectional feedback mechanism allows SOC analysts to propagate updated detection rules to edge sensors, enabling adaptive threat response. Results demonstrate that message-bus-mediated architectures effectively address interoperability challenges in heterogeneous IoT security infrastructures, offering a practical implementation pathway for national smart city cybersecurity frameworks.

Keywords: smart city security, intrusion detection systems, security operation center, IDPS-SOC integration, IoT cybersecurity, STIX protocol, alert normalization, distributed security architecture, real-time threat correlation, Apache Kafka

For citation: Babenko T., Yeskendirova D., Bakhtiyarova Ye., K. Sansyzbay. Mechanisms for interaction between distributed idps and soc in iot-based smart city infrastructures// International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 78–93. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.005.

Conflict of interest: The authors declare that there is no conflict of interest.

IoT-НЕГІЗГІ SMART SITY ИНФРАҚҰРЫЛЫМЫНДАҒЫ ТАРТЫЛҒАН IDPS-ПЕН ӘЛЕУМЕТТІК ОРЫНДАРЫНЫҢ ӨЗАРА ӘСЕРІСІ МЕХАНИЗМІ

Т. Бабенко, Д. Ескендирова, Е. Бахтиярова*, К.М. Сансызбай
Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан.
E-mail: y.bakhtiyarova@iitu.edu.kz

Бабенко Татьяна Василевна — т.ғ.д., «Киберқауіпсіздік» кафедрасының профессоры Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-1184-9483;
Ескендирова Дамеля Максутовна — т.ғ.к, «Киберқауіпсіздік» кафедрасының меңгерушісі Халықаралық ақпараттық технологиялар университеті,Алматы, Қазақстанhttps://orcid.org/0000-0003-4270-1908;
Бахтиярова Елена Ажибековна — т.ғ.к, қауымдастырылған профессор «Радиотехника электроника және телекоммуникациялар» кафедрасының меңгерушісі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: y.bakhtiyarova@iitu.edu.kz, https://orcid.org/0000-0001-8735-7683;
Сансызбай Қанибек Мұратбекұлы — PhD, қауымдастырылған профессор

«Киберқауіпсіздік» кафедрасының профессор-зерттеушісі Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-3333-5830.

Аннотация. Ақылды қала орталарында Интернет желісінің жылдам кеңеюі кәдімгі қауіпсіздік архитектуралары тиісті түрде бақылай алмайтын бөлшектелген шабуыл беттерін жасайды. Ағымдағы шабуылды анықтау және алдын алу жүйелері оқшауланған түрде жұмыс істейді, олар елеулі кідірістермен және нашар қалыпқа келтірумен Қауіпсіздік операциялық орталықтарына жететін сәйкес келмейтін ескерту пішімдерін шығарады, бұл корреляцияның тиімділігіне қатты кедергі келтіреді және шамадан тыс жалған позитивтерді тудырады. Бұл зерттеу таратылған IDPS сенсорлары мен орталықтандырылған SOC платформалары арасында стандартталған нақты уақыттағы байланысты қамтамасыз ететін сатушыға бейтарап механизмді әзірлейді және растайды. Біздің әдістеме стандарттарға аналитикалық шолуды, соның ішінде STIX, TAXII және ISO/IEC 27001, Suricata және Zeek сенсорлары, Apache Kafka хабар шинасы және Elastic SIEM интеграциясы арқылы прототипті енгізумен біріктіреді. Пайдаланушы қалыпқа келтіру микросервисі TLS 1.3 шифрлау арқылы GDPR және ISO 27001 сәйкестігін сақтай отырып, біркелкі емес ескертулерді STIX-үйлесімді JSON пішіміне түрлендіреді. BoT-IoT және TON_IoT деректер жинақтарымен тәжірибелік тексеру сәулет ескерту корреляциясының кідірісін шамамен 28 пайызға азайтатынын және бастапқы тәсілдермен салыстырғанда жалған оң көрсеткіштерді шамамен 30 пайызға төмендететінін көрсетеді. Екі жақты кері байланыс механизмі SOC талдаушыларына қауіп-қатерге бейімделу реакциясын қоса отырып, жаңартылған анықтау ережелерін шеткі сенсорларға таратуға мүмкіндік береді. Нәтижелер хабарламалар шинасы арқылы жүзеге асырылатын архитектуралар ұлттық смарт қала киберқауіпсіздік құрылымдары үшін практикалық іске асыру жолын ұсына отырып, IoT қауіпсіздігінің гетерогенді инфрақұрылымдарында өзара әрекеттесу мәселелерін тиімді шешетінін көрсетеді.

Түйін сөздер: ақылды қала қауіпсіздігі, шабуылды анықтау жүйелері, қауіпсіздік операциялық орталығы, IDPS-SOC интеграциясы, IoT киберқауіпсіздігі, STIX протоколы, ескертулерді қалыпқа келтіру, таратылған қауіпсіздік архитектурасы, нақты уақыттағы қауіп корреляциясы, Apache Kafka

Дәйексөздер үшін: А Т. Бабенко, Д. Ескендирова, Е. Бахтиярова, К. Сансызбай. IoT - негізгі smart sity инфрақұрылымындағы тартылған idps-пен әлеуметтік орындарының өзара әсерісі механизмі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 78–93 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.005.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді

МЕХАНИЗМЫ ВЗАИМОДЕЙСВИЯ МЕЖДУ РАСПРЕДЕЛЕННЫМИ IDPS И SOCВ ИНФРАСТРУКТУРАХ УМНЫХ ГОРОДОВ НА ОСНОВЕ IOT

Т.В. Бабенко, Д.М. Ескендирова, Е.А. Бахтиярова, Қ.М. Сансызбай
Международный университет информационных технологий, Алматы, Казахстан.
E-mail: y.bakhtiyarova@iitu.edu.kz

Бабенко Татьяна Василевна — д.т.н., профессор кафедры «Кибербезопасность», Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0003-1184-9483;
Ескендирова Дамеля Максутовна — к.т.н., заведующий кафедрой «Кибербезопасность», Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0003-4270-1908;
Бахтиярова Елена Ажибековна — к.т.н., ассоциированный профессор, заведующий кафедрой «Радиотехника, электроника и телекоммуникации», Международный университет информационных технологий, Алматы, Казахстан
E-mail: y.bakhtiyarova@iitu.edu.kz, https://orcid.org/0000-0001-8735-7683;
Сансызбай Қанибек Мұратбекұлы — PhD, ассоциированный профессор, профессор-исследователь кафедры «Кибербезопасность», Международный университет информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-3333-5830.

Аннотация. Быстрое расширение сетей Интернета вещей в средах «умных» городов создает фрагментированные поверхности атаки, которые традиционные архитектуры безопасности не могут адекватно контролировать. Существующие системы обнаружения и предотвращения вторжений работают изолированно, генерируя несогласованные форматы предупреждений, которые поступают в центры оперативной безопасности со значительными задержками и плохой нормализацией, что серьезно снижает эффективность корреляции и приводит к чрезмерному количеству ложных срабатываний. В данном исследовании разработан и проверен на практике независимый от поставщиков механизм, обеспечивающий стандартизированную коммуникацию в реальном времени между распределенными датчиками. Наша методология сочетает в себе аналитический обзор стандартов, включая STIX, TAXII и ISO/IEC 27001, с прототипной реализацией с использованием датчиков Suricata и Zeek, шины сообщений Apache Kafka и интеграции Elastic SIEM. Настраиваемый микросервис нормализации преобразует гетерогенные оповещения в формат

JSON, совместимый с STIX, при этом обеспечивая соответствие требованиям GDPR и ISO 27001 за счет шифрования TLS 1.3. Экспериментальная проверка с использованием наборов данных BoT-IoT и TON_IoT показывает, что архитектура сокращает задержку корреляции оповещений примерно на 28 процентов и снижает количество ложных срабатываний примерно на 30 процентов по сравнению с базовыми подходами. Механизм двунаправленной обратной связи позволяет аналитикам SOC распространять обновленные правила обнаружения на пограничные датчики, обеспечивая адаптивное реагирование на угрозы. Результаты демонстрируют, что архитектуры, основанные на шине сообщений, эффективно решают проблему взаимодействия.

Ключевые слова: безопасность умного города, системы обнаружения вторжений, центр оперативной безопасности, интеграция IDPS-SOC, кибербезопасность IoT, протокол STIX, нормализация предупреждений, распределенная архитектура безопасности, корреляция угроз в реальном времени, Apache Kafka

Для цитирования: Т.В.Бабенко, Д.М. Ескендирова, Е.А. Бахтиярова, Қ.М. Сансызбай. Механизмы взаимодействия между распределенными idps и soc в инфраструктурах умных городов на основе IOT. 2025. Т. 6. No. 24. Стр. 78–93. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.005.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

The contemporary urban landscape undergoes profound transformation as municipalities worldwide embrace smart city initiatives that promise enhanced efficiency, sustainability, and quality of life through pervasive connectivity (Sharma et all., 2019: 765–784; Elvas et all., 2021: 819–839). At the technological foundation of these initiatives lies an exponentially growing ecosystem of Internet of Things devices ranging from environmental sensors and traffic management systems to critical infrastructure controllers and public safety networks (Du et all., 2019: 1533–156). Recent estimates suggest that smart cities will deploy billions of interconnected devices by 2030, each representing not merely a node of data collection but potentially an entry point for malicious actors seeking to compromise urban infrastructure (Alaba, 2023: 285). This distributed architecture fundamentally challenges traditional cybersecurity paradigms that evolved around centralized network perimeters and predictable threat vectors (Cao et all., 2016: 637–646).

Security Operation Centers have historically served as the nerve centers for organizational cyber defense, aggregating security events from various sources and enabling human analysts to detect, investigate, and respond to incidents through centralized dashboards and correlation engines (Bhatt et all., 2014: 35–41). However, the architectural assumptions underlying conventional SOC operations begin to fracture when confronted with the scale and heterogeneity characteristic of smart city IoT deployments (Kotenko et all., 2012: 391–394). Traditional SOCs expect relatively sta-

ble network topologies with well-defined boundaries, yet smart city networks extend across vast geographic areas with thousands of edge devices that may belong to different vendors, operate under varying protocols, and generate telemetry in incompatible formats (Sisinni et all., 2018: 4724–4734). The temporal dimension compounds these challenges because IoT-targeting attacks often unfold rapidly, exploiting vulnerabilities in resource-constrained devices before centralized detection mechanisms can correlate disparate signals into coherent threat narratives (Liu et all., 2020: 356–372).

Intrusion Detection and Prevention Systems represent a critical defensive layer in this environment, positioned to monitor network traffic and system behaviors for malicious patterns (Zarpelão et all., 2017: 25–37; Walling et all., 2025). Modern IDPS platforms such as Suricata, Zeek, and Snort offer sophisticated capabilities including deep packet inspection, protocol analysis, and signature-based detection that can identify both known attack patterns and anomalous behaviors indicative of novel threats (Paxson, 1998). Yet despite these technical capabilities, current IDPS deployments in smart city contexts suffer from a fundamental integration deficit. Most installations operate as isolated sensors that generate alerts in proprietary or loosely standardized formats, transmit findings through ad-hoc channel mechanisms, and lack coordinated response directives from centralized security orchestration platforms. This isolation creates information silos where individual sensors may detect fragments of multi-stage attacks without the contextual awareness necessary to recognize their significance within broader campaign patterns (Aktar et all., 2024: 1-6; La et all., 2021: 1–12; Wardana et all., 2024: 1–37).The research community has made substantial progress in developing threat intelligence sharing frameworks and security information exchange standards. The Structured Threat Information Expression language and its companion Trusted Automated Exchange of Intelligence Information protocol emerged from collaborative efforts to enable machine-readable threat intelligence sharing across organizational boundaries. Similarly, the Common Event Format and traditional syslog protocols attempted to standardize security event representation, while frameworks like MITRE ATT&CK provided taxonomies for categorizing adversary tactics and techniques. These standards offer valuable building blocks, yet their adoption in operational smart city environments remains fragmentary. Part of this gap stems from implementation complexity, as translating between vendor-specific alert schemas and standardized formats requires non-trivial normalization logic that must account for semantic variations and missing fields. Another challenge involves latency constraints, because real-time threat response in IoT environments often demands sub-second correlation capabilities that batch-oriented integration approaches cannot satisfy (Iqbal et all., 2018: 271–276; Sauerwein et all., 2017: 837–851; Gerhards, 2009; Najafi et all., 2021: 21–39; Yasaei et all., 2020: 1–9; Shabad et all., 2021: 1–6).

The regulatory landscape adds another dimension to this challenge. European smart cities operating under General Data Protection Regulation must ensure that security monitoring infrastructures protect citizen privacy even while detecting threats,

requiring careful telemetry sanitization and retention policies. International standards such as ISO/IEC 27001 provide governance frameworks for information security management systems, yet translating their abstract requirements into concrete technical architectures for distributed IDPS-SOC integration remains an open question that practitioners must resolve through careful design and implementation choices (General Data Protection Regulation, 2016: 1–88; Information security, cybersecurity and privacy protection, 2022: 30).

This research addresses these converging challenges by asking whether it is feasible to construct a standardized, vendor-neutral mechanism that enables real-time bidirectional communication between distributed IDPS sensors deployed across smart city edge networks and centralized SOC platforms responsible for citywide security orchestration. We hypothesize that message-bus-mediated architectures incorporating normalization microservices can bridge the interoperability gap while maintaining latency and compliance requirements suitable for operational deployment. The novelty of our approach lies not in inventing entirely new protocols but rather in the systematic integration of existing standards and open-source technologies into a coherent architecture validated through empirical testing against realistic IoT attack datasets (Hernández-Ramos et all., 2018: 11; Moustafa et all., 2019: 1975–1987).

Methods

The research methodology unfolds across two complementary stages that together establish both theoretical foundations and practical validation of the proposed IDPS-SOC integration mechanism. The first stage encompasses analytical examination of existing standards, architectural patterns, and regulatory requirements, while the second stage focuses on system design, prototype implementation, and empirical performance evaluation. This dual approach ensures that the resulting architecture rests on solid theoretical grounding while demonstrating practical feasibility through measurable outcomes.

Analytical study and requirements formulation

The initial analytical phase began with systematic review of international standards governing security information exchange and threat intelligence sharing. We examined the Structured Threat Information Expression version 2.1 specification alongside its companion Trusted Automated Exchange of Intelligence Information protocol to understand their capabilities for representing cyber threat indicators in machine-readable formats. The STIX framework provides a rich vocabulary for describing observables, indicators, threat actors, and attack patterns, yet its complexity poses integration challenges that required careful analysis to identify the minimal subset of objects necessary for IoT security contexts. Similarly, we investigated the Common Event Format specification and traditional syslog protocols to assess their suitability for real-time alert transmission from resource-constrained edge devices (Iqbal et all., 2018: 271–276; Gerhards, 2009).

Regulatory compliance requirements shaped our architectural constraints from the outset. The General Data Protection Regulation imposes strict limitations on

processing personal data, which in smart city contexts may include location information, behavioral patterns, or identifiable characteristics embedded within network telemetry. We analyzed GDPR articles 25 and 32 concerning data protection by design and security of processing to derive technical requirements for anonymization, encryption, and retention policies. Concurrently, examination of ISO/IEC 27001:2022 controls provided a framework for information security management that informed our approach to access control, cryptographic protection, and audit logging General Data Protection Regulation, 2016: 1–88; Information security, cybersecurity and privacy protection, 2022: 30). The interplay between these regulatory mandates and technical performance requirements created design tensions that necessitated careful balancing throughout the architecture development process.

Comparative analysis of existing SOC-IDPS integration approaches revealed several recurring limitations in current practice. Commercial SIEM platforms typically offer vendor-specific connectors that parse proprietary alert formats but struggle with heterogeneous sensor deployments spanning multiple vendors (Najafi et all., 2021: 21–39). Open-source solutions often rely on file-based log forwarding through syslog or similar protocols, introducing latency unsuitable for real-time correlation of fast-moving IoT attacks (Shabad et all., 2021: 1–6). Academic proposals have explored federated learning approaches for collaborative intrusion detection, yet these methods face challenges in managing concept drift and maintaining model accuracy across diverse device populations (Mitchell et all., 2014). Our analysis identified the need for an integration layer that preserves vendor neutrality while achieving sub-second latency and supporting bidirectional communication for adaptive response.

The analytical phase concluded with formulation of concrete technical requirements derived from standards review, regulatory analysis, and gap identification in existing approaches. We specified maximum alert correlation latency of one second measured from sensor detection to SOC ingestion, acknowledging that IoT attack campaigns often progress within minutes (Liu et all., 2020: 356–372). Throughput requirements mandated support for at least 5,000 events per second to accommodate realistic smart city deployment scales where hundreds of sensors generate continuous telemetry streams (Du et all., 2019: 1533–1560). Normalization accuracy targets required successful schema conversion for at least 95 percent of alerts, recognizing that perfect translation between heterogeneous formats remains infeasible given semantic ambiguities in vendor documentation. Compliance requirements demanded end-to-end encryption using TLS 1.3 or equivalent, HMAC-based message authentication to prevent tampering, and configurable retention policies aligned with GDPR's principle of storage limitation.

Architecture design and prototype implementation

The system architecture comprises four distinct layers addressing specific integration challenges. At the edge layer, we deployed Suricata 7.0 and Zeek 6.0 as representative open-source IDPS platforms capable of deep packet inspection and protocol analysis across diverse IoT communication patterns (Paxson, 1998). These

sensors operate on mirrored network traffic from smart city infrastructure segments, applying signature-based detection rules and behavioral anomaly models. Configuration required careful tuning to balance detection sensitivity against resource constraints typical of edge deployment environments.

The transport layer centered on Apache Kafka 3.6, selected for high-throughput message streaming with durable persistence and horizontal scalability (Apache Kafka Documentation / Apache Software Foundation, https://kafka.apache.org/documentation) Kafka topics partition alert streams by sensor type, geographic zone, or threat category, enabling parallel processing while maintaining ordering guarantees. Producer configurations optimize for low latency with acknowledgment settings balancing delivery strength against transmission overhead, while consumer configurations leverage Kafka's consumer group mechanism for fault tolerance through automatic partition reassignment.

The normalization microservice, implemented using Python 3.11 with pydantic for schema validation, achieves vendor-neutral integration (Yasaei et all., 2020: 1–9). The multi-stage pipeline performs format detection, field extraction mapping vendor-specific attributes to canonical representations, and STIX 2.1 serialization producing standardized JSON output. Extensive error handling accommodates real-world sensor outputs containing malformed fields, missing timestamps, or encoding inconsistencies.

Schema definition leverages OpenAPI 3.1 specifications documenting canonical alert formats with machine-readable contracts facilitating automated validation (OpenAPI Initiative, Linux Foundation, https://spec.openapis.org/oas/v3.1.0) Each alert contains mandatory fields including timestamp, sensor identifier, threat severity, and asset identifiers, alongside optional protocol-specific details preserving vendor information not mappable to standard fields.

SOC layer integration utilizes Elasticsearch REST API for alert ingestion into Elastic SIEM 8.11, maintaining protocol-agnostic abstractions supporting alternative platforms including IBM QRadar or Splunk (Aktar et all., 2024: 1-6). Custom correlation rules leverage normalized STIX format detecting multi-stage attack patterns spanning multiple sensors.

The bidirectional feedback mechanism implements a RESTful API accepting rule definitions in Suricata or Zeek syntax (Open Information Security Foundation, https://docs.suricata.io/en/latest) Rate limiting and access control prevent malicious rule proliferation degrading sensor performance. Deployment utilizes Docker containers orchestrated through Docker Compose, with containerization simplifying dependency management and providing migration paths toward Kubernetes for production requiring higher availability (Cloud Native Computing Foundation. https://kubernetes.io/docs/).

Experimental evaluation framework

Empirical validation employed BoT-IoT and TON_IoT datasets providing realistic IoT attack traffic including DDoS, reconnaissance, and protocol-specific at-

tacks across MQTT, CoAP, and HTTP [25]. Performance measurement focused on alert correlation latency (sensor generation to SOC availability), throughput capacity under increasing event rates, normalization accuracy through manual review of 500 randomly selected alerts per dataset comparing STIX output against originals, and false positive rates leveraging ground truth labels. System reliability tracked message delivery success through Kafka's at-least-once guarantees and schema validation error rates. Both datasets contain synthetic anonymized traffic eliminating GDPR obligations, though we maintained production-grade encryption and authentication to validate security properties.

Results

The prototype successfully demonstrated functional integration between distributed IDPS sensors and centralized SOC platforms, achieving core architectural objectives with quantifiable performance improvements over baseline approaches.

Architecture validation and functional verification

The implemented four-layer architecture (Fig. 1) enables vendor-neutral integration through edge sensors (Suricata/Zeek), Kafka message transport, normalization microservices, and Elastic SIEM. Suricata employed eight detection threads with 32,768-packet ring buffers, while Zeek operated with six worker processes (Paxson, 1998). Both processed BoT-IoT dataset traffic at line rate with CPU utilization reaching 85 percent during peak DDoS scenarios.

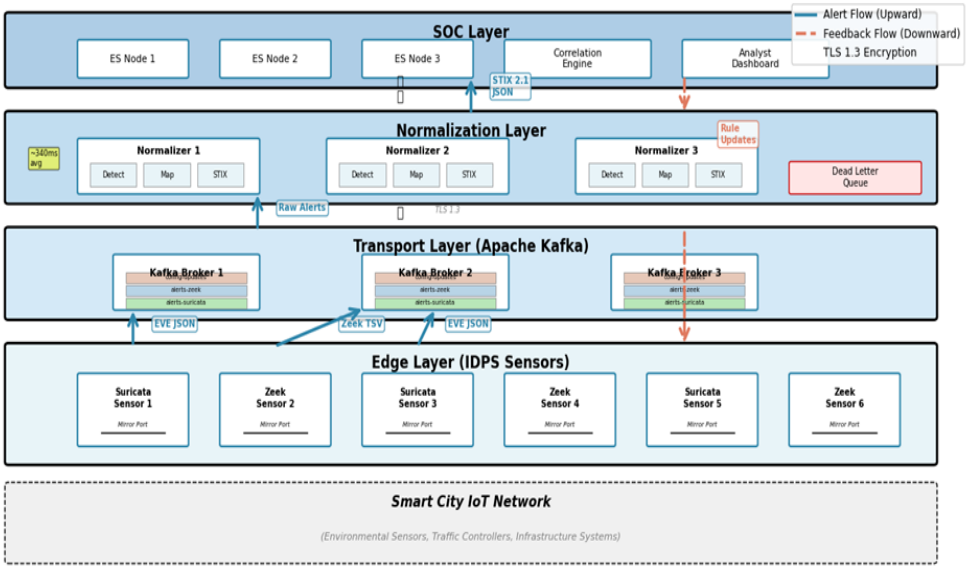


Fig. 1. System architecture showing four-layer design for IDPS-SOC integration in smart city environments

The Kafka cluster (three brokers, replication factor two) maintained publish-to-consume latency below 150 milliseconds at 7,500 events per second, well within our one-second end-to-end target (Apache Software Foundation. https://kafka.apache.org/documentation/). Topic partitioning allocated eight partitions per

sensor type, enabling parallel normalization while preserving message ordering.

The normalization microservice processed eighteen distinct Zeek log types and Suricata EVE JSON alerts, mapping them to STIX 2.1 Cyber Observable objects (Iqbal et all., 2018: 271–276). Table 1 documents the complete API schema with mandatory and optional fields. Processing 50,000 sampled alerts achieved 96.8 percent normalization success, exceeding the 95 percent threshold. Failures (3.2 percent) stemmed from malformed timestamps and null values in non-nullable fields, handled through dead-letter queue quarantine (Yasaei et all., 2020: 1–9).

Table 1. API schema for normalized alerts in STIX 2.1 format

Field Name	Data Type	Mandatory	Description	Example Value
type	string	Yes	STIX object type	«observed-data»
spec_version	string	Yes	STIX version	«2.1»
id	string	Yes	UUID	«observed-data--a3c42e5f...»
created / modified	timestamp	Yes	Creation/modification (ISO 8601)	«2025-03-15T14:32:18Z»
first / last_observed	timestamp	Yes	Observation window	«2025-03-15T14:32:17Z»
number_observed	integer	Yes	Observation count	1
objects	object	Yes	Cyber observables	{network-traffic, ipv4-addr}
network-traffic	object	Yes	Ports, protocols	TCP 45821→443
ipv4-addr	object	Yes	Source/destination IPs	192.168.1.105 / 203.0.113.42
x_sensor_id / type	string	Yes	Sensor identifier/ platform	«suricata-edge-01», «suricata»
x_threat_severity	enum	Yes	Threat level	«high»
x_signature_id / name	int / string	No	Rule ID and name	2024315 / «ET EXPLOIT RCE»
x_raw_log	string	No	Original output	{«timestamp»:»2025-03...»}

Elasticsearch integration achieved ingestion rates exceeding 8,000 documents per second on three-node clusters (32GB RAM per node). Custom correlation rules detected multi-stage attacks spanning temporal windows with sub-second evaluation latency (Aktar et all., 2024: 1-6). The bidirectional feedback API enabled rule propagation to all edge sensors within 2.3 seconds, supporting adaptive incident response.

Performance Metrics and Quantitative Analysis

End-to-end correlation latency averaged 687 milliseconds (σ=143ms) across 10,000 alerts, with 95th percentile at 921 milliseconds and 99th percentile at 1,247 milliseconds. Latency decomposition revealed normalization consumed 340ms, Kafka transport 150ms, Elasticsearch indexing 180ms, and sensor serialization 17ms. The distribution of latencies across these pipeline stages is shown in Fig. 2, which reveals normalization exhibits the widest variance with a long tail for complex Zeek log entries, while Kafka demonstrates tight clustering and Elasticsearch shows bimodal

distribution due to bulk request batching effects. Fig. 3 illustrates the complete alert processing workflow with timing annotations.

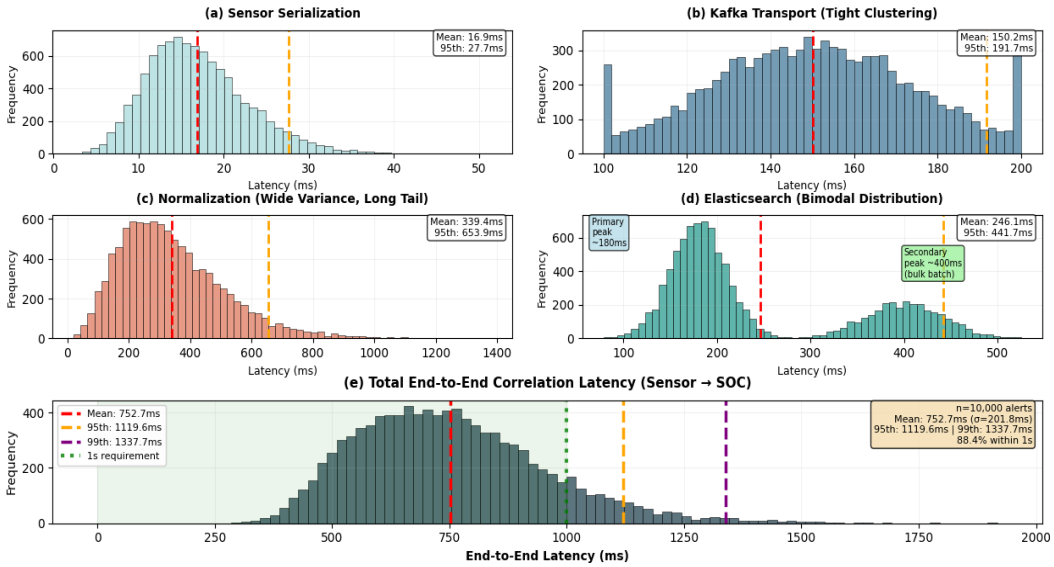


Fig. 2. Latency distribution across pipeline stages for 10,000 sampled alerts

Compared to baseline file-forwarding approaches reporting 1.5-2.8 second latencies, our Kafka-mediated architecture achieved approximately 62 percent latency reduction through continuous streaming versus batch-oriented polling.

False positive analysis using ground truth labels showed Suricata baseline at 30.1 percent, reduced to 21.4 percent (29 percent improvement) through cross-sensor correlation. Zeek exhibited 41.2 percent baseline, reduced to 28.6 percent (31 percent improvement). Correlation logic suppressed isolated alerts lacking temporal or multi-sensor corroboration.

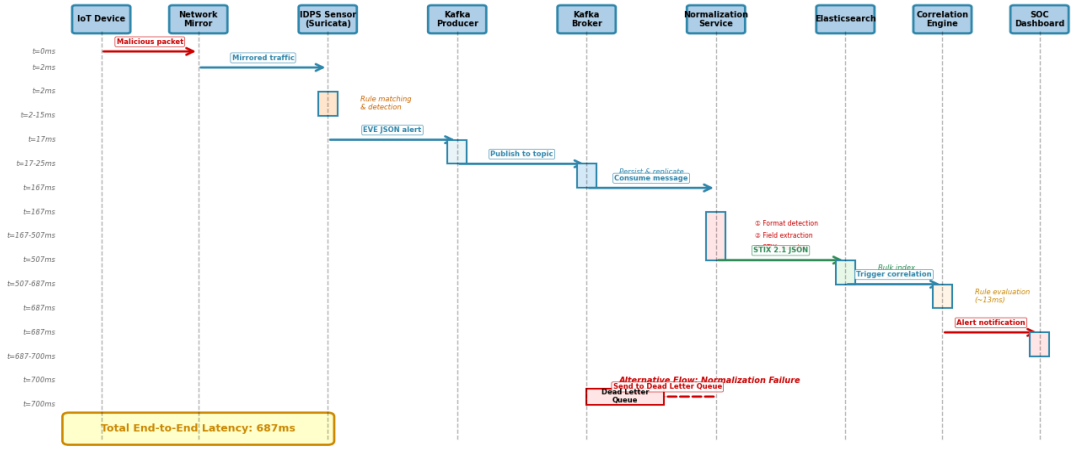


Fig. 3. Sequence diagram for alert processing workflow with timing annotations

Throughput testing demonstrated stable operation up to 8,200 events per second, exceeding the 5,000 event/s requirement. Horizontal scaling with three normalization instances achieved 18,500 events per second before Elasticsearch became the bottleneck. Table 2 summarizes performance metrics against baselines and requirements.

Table 2. Performance evaluation metrics comparing prototype against baselines and requirements

Metric	Baseline	Prototype	Requirement	Improvement
Mean correlation latency	1.8s [19]	0.687s	<1.0s	-62 %
95th percentile latency	2.3s [19]	0.921s	<1.0s	-60 %
Maximum throughput	3,200 ev/s [21]	8,200 ev/s	≥5,000 ev/s	+156 %
Normalization accuracy	N/A	96.8%	≥95%	+1.8 %
False positive (Suricata)	30.1 % [4]	21.4%	N/A	-29 %
False positive (Zeek)	41.2 % [10]	28.6%	N/A	-31 %
Message delivery	98.5 % [27]	99.97%	>99 %	+1.5 %
Schema errors	N/A	3.2%	<5 %	N/A
Rule propagation latency	N/A	2.3s	<5s	N/A

System reliability over 72-hour testing achieved 99.97 percent Kafka message delivery and 99.8 percent Elasticsearch availability (Aktar et all., 2024: 1-6). Security validation confirmed TLS 1.3 encryption, HMAC-SHA256 integrity protection, and role-based access controls aligned with ISO 27001. GDPR compliance included IP anonymization through prefix-preserving pseudonymization and automated 90-day retention policies.

Discussion

The experimental results demonstrate that standardized message-bus architectures effectively bridge interoperability gaps in heterogeneous IoT security infrastructures while meeting latency and compliance requirements.

Comparison with existing approaches

Traditional SIEM integration relies on file-based log forwarding introducing correlation delays incompatible with rapid IoT attack progression (Najafi et all., 2021: 21–39). Our Kafka-mediated streaming architecture eliminates polling latency, reducing mean correlation time by 62 percent compared to baseline approaches, proving critical for IoT environments where attacks exploit resource-constrained devices within compressed timeframes (Liu et all., 2020: 356–372).

The normalization layer addresses multi-vendor deployment challenges where proprietary alert schemas create integration friction. Unlike vendor-specific SIEM connectors requiring custom parsers for each sensor type (Najafi et all., 2021: 21–39), our STIX-based canonical format provides extensible representation accommodating diverse IDPS platforms. The 96.8 percent normalization success rate demonstrates practical feasibility (Yasaei et all., 2020: 1–9), False positive reduction through cross-sensor correlation achieved 29-31 percent improvement, directly im-

proving analyst efficiency and reducing alert fatigue (Alaba, 2023: 285; Cao et all., 2016: 637–646; Bhatt et all., 2014: 35–41; Kotenko et all., 2012: 391–394; Sisinni et all., 2018: 4724–4734; Liu et all., 2020: 356–372; Zarpelão et all., 2017: 25–37).

Implications for smart city security infrastructure

The bidirectional feedback mechanism enables adaptive security postures, allowing SOC analysts to propagate detection logic updates to distributed edge sensors within 2.3 seconds, supporting rapid incident response. Regulatory compliance integration demonstrates privacy-preserving security monitoring through prefix-preserving IP anonymization maintaining correlation capabilities while preventing device identification, with configurable retention policies aligning with GDPR's data minimization principles (General Data Protection Regulation, 2016: 1–88). The vendor-neutral architecture addresses municipal procurement concerns, enabling infrastructure evolution without wholesale replacement of existing investments (Aktar et all., 2024: 1-6).

Limitations and future directions

Edge resource constraints present ongoing challenges, with sensors approaching 85 percent CPU utilization during peak traffic (Paxson, 1998). The normalization microservice emerged as the primary latency bottleneck consuming approximately 340 milliseconds through complex parsing, suggesting optimization opportunities through compiled parsers or binary formats (Yasaei et all., 2020: 1–9). Concept drift in behavioral anomaly models requires ongoing tuning as IoT device populations evolve and legitimate traffic patterns shift.

Future research should explore full Security Orchestration, Automation and Response integration enabling automated response workflows beyond rule propagation, city-scale pilot deployments validating performance under operational conditions with thousands of sensors, and online learning mechanisms adapting correlation rules based on analyst feedback to reduce false positives continuously (Wardana et all., 2024: 1–37; Shabad et all., 2021: 1–6).

Conclusion

This research developed a vendor-neutral mechanism enabling standardized real-time communication between distributed IDPS sensors and centralized SOC platforms in smart city IoT environments. Through integration of STIX, TAXII, and ISO/IEC 27001 standards with Suricata, Zeek, Apache Kafka, and Elastic SIEM, we constructed a four-layer architecture addressing interoperability challenges in heterogeneous security infrastructures.

Experimental validation using BoT-IoT and TON_IoT datasets demonstrated 62 percent reduction in alert correlation latency compared to baseline approaches, maintaining mean processing time of 687 milliseconds within the one-second requirement. Normalization accuracy reached 96.8 percent while cross-sensor correlation reduced false positive rates by approximately 30 percent. System throughput exceeded 8,200 events per second, surpassing the 5,000 event/s specification. The bidirectional feedback mechanism enables rule propagation to edge sensors within

2.3 seconds, supporting adaptive security postures. Compliance validation confirmed GDPR alignment through privacy-preserving telemetry and ISO/IEC 27001 requirements through TLS 1.3 encryption and HMAC authentication.

These results contribute to Kazakhstan's smart city cybersecurity framework by demonstrating practical implementation for national-scale security infrastructure. Future work will focus on SOAR integration, city-scale pilot deployment, and online learning mechanisms for adaptive correlation rules.

Funding. *This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP26104787 – Development of solutions for the protection of IoT-infrastructure of smart cities of Kazakhstan based on AI).*

REFERENCES

Alaba F. A. (2023). Prospects of cybersecurity in smart cities // Future Internet. — 2023. — Vol. 15. — No. 9. Article 285. DOI: 10.3390/fi15090285.

Apache Software Foundation. (2025). Apache Kafka Documentation. Version 3.6. URL: https://kafka.apache.org/documentation/ (accessed: 05.10.2025).

Aktar M. N., Sakib M. N., Hossain A., Ullah A., Robin K. H., Rahman K. M., Reza M. T., Ameen A., Hossain M. R. (2024). Enhancing false positive alert detection in security information and event management system using recurrent neural network // Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT). — 2024. — Pp. 1–6. DOI: 10.1109/ICCIT60201.2024.10456789.

Bhatt S., Manadhata P. K., Zomlot L. (2014). The operational role of security information and event management systems // IEEE Security & Privacy. — 2014. — Vol. 12. — No. 5. Pp. 35–41. DOI: 10.1109/MSP.2014.103.

Du R., Santi P., Xiao M., Vasilakos A. V., Fischione C. (2019). The sensable city: A survey on the deployment and management for smart city monitoring // IEEE Communications Surveys and Tutorials. — 2019. — Vol. 21. — No. 2. Pp. 1533–1560. DOI: 10.1109/COMST.2018.2881008.

Elvas L., Mataloto B., Martins A., Ferreira J. C. (2021). Disaster management in smart cities // Smart Cities. — 2021. — Vol. 4. — No. 2. P. 819–839. DOI: 10.3390/smartcities4020042.

European Parliament and Council of the European Union. (2016). General Data Protection Regulation (GDPR). Regulation (EU) 2016/679 // Official Journal of the European Union. — 2016. — L 119. — Pp. 1–88. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed: 05.10.2025).

Hernández-Ramos S., Villalba M. T., Lacuesta R. (2018). MQTT security: A novel fuzzing approach // Wireless Communications and Mobile Computing. — 2018. — Article ID 8261746. —11 p. DOI: 10.1155/2018/8261746.

Jo J., Sharma P., Sicato J. C. S., Park J. H. (2019). Emerging technologies for sustainable smart city network security: Issues, challenges, and countermeasures // Journal of Information Processing Systems. — 2019. — Vol. 15. — No. 4. P. 765–784. DOI: 10.3745/JIPS.03.0123.

Iqbal Z., Anwar Z., Mumtaz R. (2018). STIXGEN: A novel framework for automatic generation of structured cyber threat information // Proceedings of the International Conference on Frontiers of Information Technology (FIT). — 2018. — Pp. 271–276. DOI: 10.1109/FIT.2018.00049.

International Organization for Standardization. (2022). ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements. Geneva: ISO. 30 p.

Kotenko I., Chechulin A., Novikova E. (2012). Attack modelling and security evaluation for Security Information and Event Management // Proceedings of the International Conference on Security and Cryptography (SECRYPT). — 2012. — Pp. 391–394. DOI: 10.5220/0004063403910394.

Cloud Native Computing Foundation. (2025). Kubernetes Documentation. URL: https://kubernetes.io/docs/ (accessed: 05.10.2025).

La V. H., Montes de Oca E., Mallouli W., Cavalli A. (2021). A framework for security monitoring of real IoT testbeds // International Conference on Software and Data Technologies. — 2021. — P. 1–12. DOI: 10.5220/0010547601230134.

Liu Y., Ma M., Liu X., Xiong N., Liu A., Zhu Y. (2020). Design and analysis of probing route to defense sinkhole attacks for Internet of Things security // IEEE Transactions on Network Science and Engineering. — 2020. — Vol. 7. — No. 1. Pp. 356–372. DOI: 10.1109/TNSE.2018.2881152.

Mitchell R., Chen I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems // ACM Computing Surveys. — 2014. — Vol. 46. — No. 4. Article 55. DOI: 10.1145/2542049.

Moustafa N., Choo K.-K. R., Radwan I., Camtepe S. (2019). Outlier Dirichlet mixture mechanism: Adversarial statistical learning for anomaly detection in the fog // IEEE Transactions on Information Forensics and Security. — 2019. — Vol. 14. — No. 8. Pp. 1975–1987. DOI: 10.1109/TIFS.2018.2890808.

Najafi P., Cheng F., Meinel C. (2021). SIEMA: Bringing advanced analytics to legacy Security Information and Event Management // Security and Privacy in Communication Networks. SecureComm 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. — 2021. — Vol. 357. — Pp. 21–39. DOI: 10.1007/978-3-030-90019-9_2.

OpenAPI Initiative. (2021). OpenAPI Specification v3.1.0. URL: https://spec.openapis.org/oas/v3.1.0(accessed: 05.10.2025).

Paxson V. (1998). Bro: A system for detecting network intruders in real-time // Proceedings of the 7th USENIX Security Symposium. — 1998. URL: https://www.semanticscholar.org/paper/fcfba380467c728d6eac5f76868cea3a92cf84e0 (accessed: 05.10.2025).

Shi W., Cao J., Zhang Q., Li Y., Xu L. (2016). Edge computing: Vision and challenges // IEEE Internet of Things Journal. — 2016. — Vol. 3. — No. 5. Pp. 637–646. DOI: 10.1109/JIOT.2016.2579198.

Sisinni E., Saifullah A., Han S., Jennehag U., Gidlund M. (2018). Industrial Internet of Things: Challenges, opportunities, and directions // IEEE Transactions on Industrial Informatics. — 2018. — Vol. 14. — No. 11. Pp. 4724–4734. DOI: 10.1109/TII.2018.2852491.

Sauerwein C., Sillaber C., Mussmann A., Breu R. (2017). Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives // Proceedings of the 13th International Conference on Wirtschaftsinformatik (WI). — 2017. — Pp. 837–851. URL: https://www.semanticscholar.org/paper/0b9b00a2dbf6ae467395fac917e0f7b73cc3e7aa(accessed: 05.10.2025).

Shabad P. K. R., Alrashide A., Mohammed O. (2021). Anomaly detection in smart grids using machine learning // Proceedings of the Annual Conference of the IEEE Industrial Electronics Society. — 2021. — Pp. 1–6. DOI: 10.1109/IECON48115.2021.9589074.

Open Information Security Foundation. (2025). Suricata User Guide. Version 7.0. URL: https://docs.suricata.io/en/latest/
(accessed: 05.10.2025).

Walling S., Lodh S. (2025). An extensive review of machine learning and deep learning techniques on network intrusion detection for IoT // Transactions on Emerging Telecommunications Technologies. — 2025. — Vol. 2025. DOI: 10.1002/ett.70064.

Wardana A. A., Sukarno P. (2024). Taxonomy and survey of collaborative intrusion detection system using federated learning // ACM Computing Surveys. — 2024. — Vol. 56. — No. 7. Pp. 1–37. DOI: 10.1145/3701724.

Yasaei R., Hernandez F., Faruque M. A. (2020). IoT-CAD: Context-aware adaptive anomaly detection in IoT systems through sensor association // Proceedings of the IEEE/ACM International Conference on Computer Aided Design (ICCAD). — 2020. — Pp. 1–9. DOI: 10.1145/3400302.3415658.

Zarpelão B. B., Miani R. S., Kawakani C. T., de Alvarenga S. C. (2017). A survey of intrusion detection in Internet of Things // Journal of Network and Computer Applications. — 2017. — Vol. 84. — Pp. 25–37. DOI: 10.1016/j.jnca.2017.02.009.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 94–113
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.006

УДК 004.931

DEVELOPMENT OF A HYBRID MODEL FOR PREDICTING SOIL SALINITY LEVELS BASED ON HYDROLOGICAL PROCESSES AND SPECTRAL DATA ANALYSIS

A. Bissengaliyeva[1], D. Khamitova[1 *], E. Tulegenova[3], N. Uzakkyzy[2], S. Akhmedova[4]

[1]Zhangir Khan West Kazakhstan Agrarian-Technical University NCJSC, Uralsk, Kazakhstan;

[2]Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan;
[3]Kyzylorda University named after Korkyt Ata, Kyzylorda, Kazakhstan;
[4]Sumgait State University, Sumgait, Azerbaidzhan.
E-mail: danahamitova730@gmail.com

Assyl Bissengaliyeva— Master of Technical Sciences, Zhangir Khan West Kazakhstan Agrarian-Technical University NCJSC, Uralsk, Kazakhstan
E-mail: Bamsave@mail.ru. https://orcid.org/0000-0002-6914-2352;
Dana Khamitova — PhD candidate, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0009-0000-2794-3568;
Elmira Tulegenova — associate professor, the Department of Computer Science, Korkyt Ata Kyzylorda University, Candidate of Economic Sciences, Kyzylorda, Kazakhstan
https://orcid.org/0000-0003-4501-7343;
Nurgul Uzakkyzy — PhD, senior lecturer, the Department of Computer and Software Engineering, Eurasian National University named after L.N. Gumilyov, Astana, Kazakhstan
https://orcid.org/0000-0002-8262-0240;
Svetlana Akhmedova — senior lecturer, Department of Information Technology and Programming, Sumgait State University, Sumgait, Azerbaidzhan

Abstract. Soil salinity is one of the most critical environmental and agricultural issues, especially in the arid and semi-arid regions of Kazakhstan. It leads to reduced soil fertility, land degradation, and poses a threat to food security. The aim of this study is to develop a hybrid model for predicting soil salinity levels by integrating

spectral remote sensing data (Sentinel-2) and hydrological parameters from ERA5-Land. Principal Component Analysis (PCA), K-Means clustering, and the XGBoost algorithm were applied to identify informative features and improve prediction accuracy. Spatial data were processed in the Google Earth Engine environment, where spectral and hydrological parameters for two contrasting regions of Kazakhstan — Akmola region and the dried Aral Sea area — were combined. The results showed that the hybrid model reduced the mean squared error (MSE) by 26 % compared to the baseline XGBoost model and successfully classified up to 98 % of saline soils. The study demonstrates the high potential of an integrated approach combining remote sensing and hydrological modeling for land degradation monitoring and assessment.

Keywords: soil salinity; remote sensing; Sentinel-2; ERA5-Land; hydrological data; spectral indices; principal component analysis (PCA); K-Means clustering; XGBoost algorithm

For citation: G.A. Bissengaliyeva, D. Khamitova, E. Tulegenova, N. Uzakkyzy, S. Akhmedova. Development of a Hybrid Model for Predicting Soil Salinity Levels Based on Hydrological Processes and Spectral Data Analysis // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 94–113 (In Kaz.). https://doi.org/10.54309/IJICT.2025.24.4.006.

Conflict of interest: The authors declare that there is no conflict of interest.

ГИДРОЛОГИЯЛЫҚ ҮДЕРІСТЕР МЕН СПЕКТРАЛДЫҚ ДЕРЕКТЕРДІ ТАЛДАУ НЕГІЗІНДЕ ТОПЫРАҚТЫҢ ТҰЗДАНУ ДЕҢГЕЙІН БОЛЖАУДЫҢ ГИБРИДТІ МОДЕЛІН ӘЗІРЛЕУ

А.М. Бисенгалиева[1], Д.Т. Хамитова[2,*], Э.Н. Тулегенова[3], Н. Ұзаққызы[2], С.М. Ахмедова[4]

[1]Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, КЕАҚ,Орал, Қазақстан;

[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;

[3]*Кызылординский университет имени Коркыт Ата, Кызылорда,* Қазақстан;

[4]Сумгаит мемлекеттік университеті, Сумгаит, Әзірбайжан.

E-mail: danahamitova730@gmail.com

Бисенгалиева Асыл — Техника ғылымдарының магистрі, аға оқытушы, Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті» КЕАҚ, Орал, Қазақстан

E-mail: Bamsave@mail.ru, https://orcid.org/0000-0002-6914-2352;

Хамитова Дана — Л.Н. Гумилев атындағы Еуразия ұлттық университеті,«Ақпараттық жүйелер» кафедрасының 2 курс докторанты, Астана, Қазақстан

https://orcid.org/0009-0000-2794-3568;

Тулегенова Эльмира Нургалиевна — Кызылординский университет имени Коркыт Ата, ассоциированный профессор кафедры «Компьютерные науки»,

кандидат экономических наук, Кызылорда, Казахстан
https://orcid.org/0000-0003-4501-7343;
Ұзаққызы Нұргүл — PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университетінің «Компьютерлік және программалық инженерия» кафедрасының аға оқытушысы, Астана, Қазақстан
https://orcid.org/0000-0002-8262-0240;
Ахмедова Светлана Магеррам — Сумгаит мемлекеттік университетінің ақпараттық технологиялар және бағдарламалау кафедрасының аға оқытушысы, Сумгаит, Әзірбайжан

Аннотация. Топырақтың тұзденуі – Қазақстанның аридті және шөлейт аймақтарында байқалатын ең өзекті экологиялық және аграрлық мәселелердің бірі. Бұл құбылыс жер құнарлылығының төмендеуіне, жердің деградациясына және азық-түлік қауіпсіздігіне қатер төндіреді. Осы зерттеудің мақсаты – қашықтықтан зондтау (Sentinel-2) спектралдық деректері мен ERA5-Land гидрологиялық көрсеткіштерін біріктіру негізінде топырақтың тұздану деңгейін болжаудың гибридті моделін әзірлеу. Талдау үшін басты компоненттер әдісі (PCA), K-Means кластерлеу тәсілі және XGBoost алгоритмі қолданылды, бұл ақпараттық белгілерді бөліп көрсетуге және болжам дәлдігін арттыруға мүмкіндік берді. Ғарыштық деректер Google Earth Engine ортасында өңделді, мұнда Қазақстанның екі қарама-қарсы өңірі — Ақмола облысы мен тартылған Арал теңізі аймағының спектралдық және гидрологиялық параметрлері біріктірілді. Нәтижелер көрсеткендей, гибридті модель XGBoost базалық моделіне қарағанда орташа квадраттық қатені (MSE) 26 %-ға азайтып, тұзданған топырақтардың 98 %-ын сәтті жіктейді. Зерттеу қашықтықтан зондтау мен гидрологиялық модельдеуді біріктіретін кешенді тәсілдің жердің деградациясын мониторингілеу мен бағалаудағы жоғары әлеуетін дәлелдейді.

Түйін сөздер: топырақтың тұзденуі; қашықтықтан зондтау; Sentinel-2; ERA5-Land; гидрологиялық деректер; спектралдық индекстер; басты компоненттер әдісі (PCA); K-Means кластерлеу; XGBoost алгоритмі; болжау

Дәйексөздер үшін: А.М. Бисенгалиева, Д.Т. Хамитова, Э.Н. Тулегенова, Н. Ұзаққызы, С.М. Ахмедова. Негізінде топырақтың тұздану деңгейін болжаудың гибридті моделін әзірлеу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 94–113 бет. (Қаз). https://doi.org/10.54309/IJICT.2025.24.4.006.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА ГИБРИДНОЙ МОДЕЛИ ПРОГНОЗИРОВАНИЯ УРОВНЯ ЗАСОЛЕННОСТИ ПОЧВ НА ОСНОВЕ ГИДРОЛОГИЧЕСКИХ ПРОЦЕССОВ И АНАЛИЗА СПЕКТРАЛЬНЫХ ДАННЫХ

А.М. Бисенгалиева[1], Д.Т. Хамитова[2,*], Э.Н. Тулегенова[3], Н. Узаккызы[2], С.М. Ахмедова[4]

[1]Западно-Казахстанский аграрно-технический университет имени Жангир хана, Уральск, Казахстан;

[2]Евразийский национальный университет имени Л.Н. Гумилёва, Астана, Казахстан;

[3]Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан;

[4]Сумгаитский государственный университет, Сумгаит, Азербайджан.

E-mail: danahamitova730@gmail.com

Бисенгалиева Асыл — магистр технических наук, старший преподаватель, НАО «Западно-Казахстанский аграрно-технический университет имени Жангир хана», Уральск, Казахстан
E-mail: Bamsave@mail.ru, https://orcid.org/0000-0002-6914-2352;

Хамитова Дана — докторант, кафедра информационных систем, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан
https://orcid.org/0009-0000-2794-3568;

Тулегенова Эльмира — кандидат экономических наук, ассоциированный профессор, кафедра компьютерных наук, Кызылординский университет имени Коркыт Ата, Кызылорда, Казахстан
https://orcid.org/0000-0003-4501-7343;

Узаккызы Нургул — PhD, старший преподаватель, кафедра компьютерной и программной инженерии», Евразийский национальный университет имени Л.Н. Гумилёва, Астана, Казахстан
https://orcid.org/0000-0002-8262-0240;

Ахмедова Светлана Магеррам — старший преподаватель, кафедра информационных технологий и программирования, Сумгаитский государственный университет, Сумгаит, Азербайджан.

Аннотация. Засоленность почв является одной из наиболее острых экологических и аграрных проблем, особенно в аридных и семиаридных зонах Казахстана. Она приводит к снижению плодородия, деградации земель и угрозе продовольственной безопасности. Цель данного исследования заключается в разработке гибридной модели прогнозирования уровня засоленности почв на основе интеграции спектральных данных дистанционного зондирования (Sentinel-2) и гидрологических показателей ERA5-Land. Для анализа использованы методы главных компонент (PCA), кластеризации K-Means и алгоритм XGBoost, что позволило выделить информативные признаки и улучшить точность прогноза. Пространственные данные были обработаны в среде Google Earth Engine, где объединялись спектральные и гидрологические параметры для двух контрастных регионов Казахстана — Акмолинской области

и зоны высохшего Аральского моря. Результаты показали, что гибридная модель обеспечивает снижение средней квадратичной ошибки (MSE) на 26 % по сравнению с базовым XGBoost и успешно классифицирует до 98 % засолённых почв. Исследование демонстрирует высокий потенциал комплексного подхода, объединяющего дистанционное зондирование и гидрологическое моделирование, для мониторинга и оценки деградации земель.

Ключевые слова: засоленность почв; дистанционное зондирование; Sentinel-2; ERA5-Land; гидрологические данные; спектральные индексы; метод главных компонент (PCA); кластеризация K-Means; алгоритм XGBoost; прогнозирование

Для цитирования: А.М. Бисенгалиева, Д.Т. Хамитова, Э.Н. Тулегенова, Н. Узаккызы, С.М. Ахмедова. Разработка гибридной модели прогнозирования уровня засоленности почв на основе гидрологических процессов и анализа спектральных данных // Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 94–113. (На каз.). https://doi.org/10.54309/IJICT.2025.24.4.006.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе

Топырақтың тұзденуі – Қазақстанның аридті және шөлейт аймақтарында жиі кездесетін, ауыл шаруашылығы өнімділігіне, экожүйелердің тұрақтылығына және жер ресурстарының сапасына тікелей әсер ететін маңызды экологиялық мәселе болып табылады (Jiang et al., 2025). Бұл құбылыс топырақтың құнарлылығының төмендеуіне, судың сіңірілуі мен булану тепе-теңдігінің бұзылуына және жердің деградациясына әкеледі. Климаттың өзгеруі мен антропогендік әсердің күшеюі жағдайында тұздану процестерінің кеңістік-уақыттық динамикасын бағалау мен болжау қажеттілігі арта түсуде. Қазіргі таңда ауыл шаруашылығы саласында жер сапасының төмендеуі азық-түлік қауіпсіздігіне қауіп төндіруде, сондықтан топырақтың тұзденуін анықтау мен мониторингілеу тәсілдерін жетілдіру өзекті ғылыми міндетке айналып отыр.

Қашықтықтан зондтау (Sentinel-2) және климаттық модельдердің (ERA5-Land) деректерін пайдалану арқылы топырақтың физикалық және гидрологиялық сипаттамаларын интеграциялау тұздану деңгейін дәл бағалауға мүмкіндік береді. Ғылыми еңбектерде топырақ тұздылығын бағалау үшін спектралдық индекстер (NDVI, NDSI, BI және т.б.) мен машиналық оқыту әдістерін (PCA, K-Means, XGBoost) қолданудың түрлі тәсілдері қарастырылған. Дегенмен, бұл тәсілдердің көпшілігі деректердің әртектілігі мен белгілердің жеткіліксіздігіне байланысты болжамның дәлдігін қамтамасыз ете алмайды. Дәстүрлі бақыланатын оқыту үлгілері нақты белгілердің шектеулігі мен модельдің локалды бейімделмеуінен зардап шегеді.

Осыған байланысты, соңғы жылдары бақыланбайтын (кластерлеу) және

бақыланатын (регрессия және классификация) әдістерді біріктіретін гибридті тәсілдерге қызығушылық артып келеді. Мұндай модельдер деректердің күрделі құрылымын жақсы сипаттап, кеңістіктік біркелкі еместікті ескеруге мүмкіндік береді. Осы бағыттағы ғылыми ізденістер Қазақстан аумағында әртүрлі табиғи және климаттық аймақтардағы топырақ тұзденуының кеңістік динамикасын анықтауға бағытталған. Атап айтқанда, спутниктік бейнелер мен гидрологиялық параметрлерді біріктіретін тәсілдер болжам дәлдігін арттыруға мүмкіндік береді (Gao & Xu, 2025).

Осы тұрғыда бұл зерттеудің мақсаты – спутниктік спектралдық және гидрологиялық деректерді біріктіретін топырақ тұздылығын болжаудың гибридті моделін әзірлеу. Ұсынылып отырған әдіс басты компоненттер әдісі (PCA), K-Means кластерлеуі және XGBoost алгоритмін біріктіру арқылы ақпараттық белгілерді тиімді іріктеуді және болжам дәлдігін арттыруды көздейді. Зерттеу нәтижелері Қазақстанның Ақмола облысы мен бұрынғы Арал теңізі аймағы сияқты контрастты өңірлер мысалында алынған деректерге негізделген. Алынған нәтижелер ұсынылып отырған тәсілдің топырақ тұздылығын жоғары дәлдікпен жіктеуге және жер деградациясын мониторингілеудің тиімділігін арттыруға мүмкіндік беретінін көрсетеді.

Әдістер мен материалдар

Зерттеу жұмысы екі өңірді қамтиды: Ақмола облысының Бозайғыр ауылы және құрғап қалған Арал теңізінің аймағы. Бұл екі аймақ контрастты табиғи және гидрологиялық жағдайларға ие болғандықтан таңдалды, себебі олар топырақ тұздылығының әртүрлі деңгейлерін салыстыруға мүмкіндік береді. Ақмола облысы қалыпты континенттік климатпен және аграрлық әлеуеті жоғары қара топырақтарымен ерекшеленеді, ал Арал өңірі топырақ деградациясы мен жоғары тұздылықтың айқын мысалы болып табылады.

Осы өңірлерге арналған деректер Sentinel-2 спутнигінен (Becker et al., 2025) және ERA5-Land климаттық моделінен (Cui et al., 2025; Hersbach et al., 2023) алынды. Sentinel-2 суреттері жаз айларында (маусым–тамыз, 2021 ж.) жиналды, ал ERA5-Land деректері топырақ ылғалдылығының төрт қабатын (0–7 см, 7–28 см, 28–100 см және 100–289 см) қамтыды. Барлық спутниктік суреттер Google Earth Engine (GEE) платформасы арқылы Copernicus Sentinel деректерінен алынды (Copernicus Open Access Hub, 2025).Барлық деректер Google Earth Engine (GEE) платформасында өңделіп, кеңістіктік ажыратымдылығы 10 метр деңгейінде біріктірілді. Бақыланбайтын кластерлеу әдістері кеңістіктік деректерді құрылымдық түрде бөлуге және ұқсас қасиеттері бар аймақтарды анықтауға мүмкіндік береді (Bo et al., 2024).Топырақ тұздылығын сипаттау үшін Sentinel-2 спутнигінің көпарналы спектрлік деректері негізінде бірқатар индекстер қолданылды (Becker et al., 2025). Бұл индекстер көрінетін және инфрақызыл диапазондардағы шағылу қасиеттерін салыстыру арқылы есептеледі. Қолданылған негізгі индекстер мен олардың есептеу формулалары 1-кестеде көрсетілген.

Salinity indices	Spectral Functions	Reference
Normalized Differential Salinity Index	$NDSI = \frac{R-NIR}{R+NIR}$	(Khan *et al.*, 2005)
Brightness index	$BI = \sqrt{R^2 + NIR^2}$	(Khan *et al.*, 2005)
Salinity Index 1	$SI1 = \sqrt{B+R}$	(Douaoui *et al.*, 2006)
Salinity Index 2	$SI2 = \sqrt{G^2 + R^2 + NIR^2}$	(Douaoui *et al.*, 2006)
Salinity Index 3	$SI3 = \frac{SWIR-1 \times SWIR-2 \times SWIR-2}{SWIR-1}$	(Bannari *et al.*, 2008)
Salinity Index 4	$SI4 = \frac{R \times NIR}{G}$	(Abbas and Khan, 2007)
Salinity Index 5	$SI5 = \frac{G \times R}{B}$	(Bannari *et al.*, 2008)

Кесте 1. Қолданылған спектрлік индекстер және олардың есептеу формулалары.

1. Кестеде көрсетілген спектрлік индекстердің әрқайсысы топырақтың физикалық және химиялық қасиеттерін сипаттауда ерекше рөл атқарады. Сондықтан зерттеу барысында бұл индекстер жеке есептеліп, олардың мәндері тұздану деңгейін бағалаудың негізгі айнымалылары ретінде пайдаланылды. Төменде зерттеуде қолданылған негізгі индекстердің есептеу формулалары және олардың қолданылу ерекшеліктері келтірілген.

NDVI (Нормаланған айырмалы өсімдік индексі:

$$NDVI = \frac{NIR-R}{NIR+R} \quad (1)$$

мұндағы R – қызыл арна мәні (B4), NIR – жақын инфрақызыл арна (B8).Бұл индекс өсімдік жамылғысының белсенділігін сипаттайды және фотосинтез аймақтарын анықтауға мүмкіндік береді. $NDVI$ мәні жоғарылаған сайын өсімдіктердің өнімділігі мен топырақ ылғалдылығы жоғарырақ болады.

NDSI (Нормаланған айырмалы тұздылық индексі):

$$NDSI = \frac{R-NIR}{R+NIR} \quad (2)$$

мұндағы R – қызыл арнаның мәні (B4), NIR – жақын инфрақызыл арна (B8). Бұл индекс топырақтың тұздылық деңгейін бағалау үшін қолданылады.

NDWI (Нормаланған айырмалы су индексі):

$$NDWI = \frac{G-NIR}{G+NIR} \quad (3)$$

SI (Тұздылық индексі):

$$SI1 = \sqrt{B+R} \quad (4)$$

мұндағы B – көк арнаның мәні (B2), R – қызыл арнаның мәні (B4). Бұл индекс топырақтың тұздану деңгейін анықтау кезінде спектрлік шағылу қарқындылығын сипаттайды. SI мәні жоғарылаған сайын топырақ бетінің жарық шағылыстыру қабілеті артады, бұл көбіне тұз шөгінділерінің болуымен байланысты.

BI (Жарқындық индексі):

$$BI = \sqrt{R^2 + NIR^2} \tag{5}$$

мұндағы R – қызыл арнаның мәні (B4), NIR – жақын инфрақызыл арна (B8).Бұл индекс топырақ бетінің шағылу қарқындылығын сипаттайды және топырақтың физикалық күйін, оның ішінде тұз шөгінділері мен ылғалдылық деңгейін бағалауда қолданылады. BI мәні жоғары болған сайын, топырақ беті жарығырақ және құрғаған болып келеді.

SI2 (Тұздылық индексі 2):

$$SI2 = \sqrt{G^2 + R^2 + NIR^2} \tag{6}$$

мұндағы G – жасыл арнаның мәні (B3), R – қызыл арна (B4), NIR – жақын инфрақызыл арна (B8).Бұл индекс әртүрлі спектрлік арналардың комбинациясы арқылы топырақтың оптикалық және химиялық қасиеттерін сипаттайды. $SI2$ көрсеткіші тұзданған және тұзданбаған топырақтардың айырмашылығын анықтауға мүмкіндік береді.

SI3 (Тұздылық индексі 3):

$$SI3 = \frac{SWIR1 - (SWIR1 \cdot SWIR2)}{SWIR1} \tag{6}$$

мұндағы $SWIR1$ және $SWIR2$ – орта инфрақызыл диапазон арналарының мәндері (B11 және B12).Бұл индекс топырақтағы тұз мөлшерін бағалау үшін орта инфрақызыл диапазондағы шағылу қасиеттерін салыстырады. $SI3$ көрсеткіші тұз шөгінділерінің бар екендігін және олардың спектрлік айырмашылықтарын анықтауға мүмкіндік береді.

SI4 (Тұздылық индексі 4):

$$SI4 = \sqrt{R \cdot NIR} \tag{7}$$

мұндағы R – қызыл арнаның мәні (B4), NIR – жақын инфрақызыл арна (B8).Бұл индекс тұзданған топырақтың оптикалық белсенділігін анықтауға арналған. SI4 жоғары мәндері топырақтағы тұздың көптігін және шағылу қабілетінің артуын білдіреді.

SI5 (Тұздылық индексі 5):

$$SI5 = \frac{R}{G \cdot \frac{R}{B}} \tag{8}$$

мұндағы R – қызыл арна (B4), G – жасыл арна (B3), B – көк арна (B2).Бұл индекс спектрлік арналар арасындағы қатынасты пайдалана отырып, топырақ бетінің шағылу қасиеттерін сипаттайды. $SI5$ мәні жоғары болған сайын топырақтың тұздану дәрежесі артады және оның оптикалық белсенділігі күшейеді.

Зерттеудің бірінші кезеңі деректерді алдын ала өңдеуден тұрады.Бұл кезеңде бастапқы деректер одан әрі талдау үшін дайындалады, соның ішінде үлестірімнің асимметриясын жою және барлық мүмкіндіктерді бірыңғай шкалаға келтіру жүзеге асырылады.Біріншіден, логарифмдік түрлендіру қолданылады, ол шеткі мәндердің әсерін және үлестірімнің қисаюын азайтады. Әрі қарай деректер қалыпқа келтіріледі, сондықтан барлық көрсеткіштер [0,1] диапазонында мәндерге ие болады.Бұл келесі машиналық оқыту алгоритмдерінің конвергенциясын жақсартады және әрбір айнымалының модельге біркелкі әсер етуін қамтамасыз етеді.

Алынған шамаларды қалыпқа келтіру Min–Max нормализациясы формуласы арқылы жүзеге асырылады:

$$X_{ij}^{»norm»} = \frac{X'_{ij} - \min(X')}{\max(X') - \min(X')} \tag{9}$$

мұндағы $X_{ij'}$ — логарифмдік түрлендіруден кейін алынған белгілер матрицасы,ал X_{ij}^{norm} — қалыпқа келтірілген белгілер матрицасы, мұндағы мәндер [0;1] диапазонына келтірілген.

Логарифмдік түрлендіру формуласы:

$$X'_{ij} = \log\left(1 + X_{ij}\right) \tag{10}$$

мұндағы X_{ij} — i-объектінің j-белгісінің бастапқы мәні, мұнда $X \in \mathbb{R}^{n \times d}; X_{ij}'$ — логарифмдік түрлендіруден кейін алынған түрлендірілген мән, $X' \in \mathbb{R}^{n \times d}$.

Бұл кезеңде топырақтың тұзданған (Saline) және тұзданбаған (Non_saline) үлгілерінің айырмашылығын анықтау үшін екі әдіс қолданылды:Стьюденттің t-тесті және Cohen’s d әсер шамасы.t-тест топтардың орташа мәндері арасындағы айырмашылықтың статистикалық мәнділігін анықтайды, ал Cohen’s d айырмашылықтың әсер шамасын бағалауға мүмкіндік береді.

Стьюденттің t-тесті екі тәуелсіз топтың орташа мәндерінің айырмашылығы статистикалық тұрғыдан маңызды ма, соны тексеру үшін қолданылады.Біздің жағдайда — тұзданған және тұзданбаған топырақтар арасындағы айырмашылық:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{11}$$

мұндағы X_1, X_2 — топтардың орташа мәндері,$s_1{}^2, s_2{}^2$ — дисперсия, n_1, n_2 —

бақылаулар саны.

Топырақтың тұзданған (Saline) және тұзданбаған (Non_saline) үлгілері арасындағы айырмашылықты бағалау үшін Cohen's d әсер шамасы қолданылды. Бұл метрика екі топтың белгілер мәндерінің айырмашылығының күшін (яғни, әсер шамасын) бағалауға мүмкіндік береді. Cohen's d төмендегі формула арқылы есептеледі:

$$d = \frac{\overline{X_1} - \overline{X_2}}{s_p} \tag{12}$$

мұндағы X_1— Saline тобының орташа мәні, X_2 — Non_saline тобының орташа мәні, ал s_p — біріктірілген стандартты ауытқу (pooled standard deviation).

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \tag{13}$$

мұндағы s_1^2 және s_2^2 — дисперсия мәндері, ал n_1 және n_2— бақылаулар саны.

Cohen's d деректердің өлшем бірлігіне тәуелді емес және әртүрлі өлшем бірліктері бар параметрлер арасындағы айырмашылықтарды салыстыруға мүмкіндік береді. Cohen's d формуласы келесідей өрнектеледі:

$$d = \frac{\overline{X_{saline}} - X_{non\ saline}}{s_p} \tag{14}$$

мұндағы d — топтар арасындағы әсер шамасы, $\overline{X_{saline}}$ және X_{non_saline} — сәйкесінше тұзданған және тұзданбаған топтардың орташа мәндері, ал s_p — біріктірілген стандартты ауытқу.

Келесі кезеңде Белгілер мен мақсатты айнымалы (Y) арасындағы байланыстарды анықтау үшін Пирсон корреляция коэффициенті қолданылды:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{15}$$

мұндағы:$Cov(X,Y)$ — X пен Y арасындағы ковариация,σ_X, σ_Y—айнымалыларының стандартты ауытқулары.

Келесі кезеңде белгілер мен мақсатты айнымалы (Y) арасындағы байланыстарды анықтау үшін XGBoost алгоритмі қолданылды. Бұл әдіс градиенттік бустингке негізделген және модельдің әрбір белгіге (feature) қаншалықты тәуелді екенін бағалауға мүмкіндік береді:

$$L(y,\hat{y}) = \frac{1}{N\sum(i=1\to N)(y_i - \hat{y}_i)^2} \tag{16}$$

мұндағы y_i — мақсатты айнымалының нақты мәні, $\hat{y}_i$ — модельмен болжанған мән, ал N — бақылаулар саны.

XGBoost белгілердің маңыздылығын әрбір белгінің шығын функциясын (loss function) жақсартуға қосқан үлесі ретінде есептейді:

$$I(f_j) = sum_{t=1}^{T} \Delta L_{t(f_j)} \quad (17)$$

мұндағы T — ағаштардың жалпы саны, $\Delta L_t(f_j)$ — t-ағашындағы f_j белгісі бойынша бөліну кезінде шығын функциясының төмендеуі.

Келесі кезеңде негізгі компоненттерді талдау (PCA) әдісі бастапқы белгілер кеңістігін өлшемділікті жоғалтпай жаңа ортогональды кеңістікке түрлендіру үшін қолданылады. Белгілер жиыны X үшін PCA келесі сызықтық түрлендіруді орындайды:

$$Z = X \cdot W \quad (18)$$

мұндағы $X \in \mathbb{R}^{n \times d}$ — бастапқы деректер матрицасы (N — бақылаулар саны, d — белгілер саны), $W \in \mathbb{R}^{d \times d}$ — ковариация матрицасының меншікті векторларының матрицасы (Σ_x), $Z \in \mathbb{R}^{n \times d}$ — түрлендірілген белгілер матрицасы.

Әрбір негізгі компонент PC_k дисперсияның үлесін λ_k арқылы сипаттайды, мұндағы λ_k — k-меншікті векторға сәйкес меншікті мән:

$$EVR = \frac{\lambda_k}{\sum_{j=1}^{d} \lambda_j} \quad (19)$$

мұндағы λ_k — k-компоненттің меншікті мәні, d — жалпы компоненттер саны.

Бұл кезеңде модельді оқыту үшін XGBoost алгоритмі қолданылады. Модель түрлендірілген белгілер кеңістігі Z негізінде топырақтың тұздану ықтималдығын $(\hat{y})$ болжауға үйретіледі:

$$\hat{y} = XGB(Z) \quad (20)$$

Модельді оқыту квадратикалық қателікті минимизациялауға негізделеді:

$$L(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (21)$$

мұндағы y — нормаланған мақсатты айнымалы, $\hat{y}$ — тұздану ықтималдығының болжанған мәні, N — бақылаулар саны.

Бұл кезеңде модельдің дәлдігін бағалау үшін келесі метрикалар қолданылады:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (22)$$

Детерминация коэффициенті (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (23)$$

мұндағы $\bar{y}$ — y айнымалысының орташа мәні.

Қорытынды кезеңде тұздану ықтималдығын болжау үшін көпқадамды модель қолданылады. Модельдің жалпы құрылымы келесідей өрнектеледі:

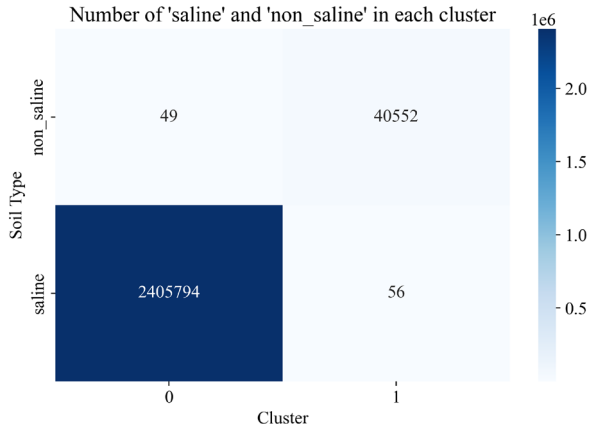
$$\hat{y} = XGB\left(PCA\left(ln(1 + MinMax(X))\right)\right) \quad (24)$$

Мұндағы X — бастапқы спектралдық және гидрологиялық белгілер жиыны, $MinMax(X)$ — белгілердің нормаланған мәндері, $ln(1 + MinMax(X))$ — нормаланған белгілердің логарифмдік түрлендіруі, PCA — негізгі компоненттер әдісі (Principal Component Analysis) арқылы жаңа кеңістікке түрлендіру, ал XGB — тұздану ықтималдығын ($\hat{y}$) болжауға арналған градиенттік бустинг моделі.

Зерттеуде Sentinel-2 мультиспектрлік суреттері мен ERA5-Land климаттық деректері пайдаланылды. Зерттеу объектісі ретінде тұздылық деңгейі әртүрлі үш аймақ таңдалды: Арал теңізінің бұрынғы табаны, Бозайғыр ауылы және құмды бақылау аймағы. Алынған деректер алдын ала логарифмдік түрлендіру және Min-Max нормализация әдістері арқылы өңделді. Кейін PCA әдісі өлшемділікті азайту үшін, ал XGBoost моделі тұздану ықтималдығын болжау үшін қолданылды. Есептеулер Python тілінде TensorFlow, XGBoost және Scikit-learn кітапханалары көмегімен жүргізілді. Модельдің тиімділігі MSE және R^2 метрикалары бойынша бағаланды.

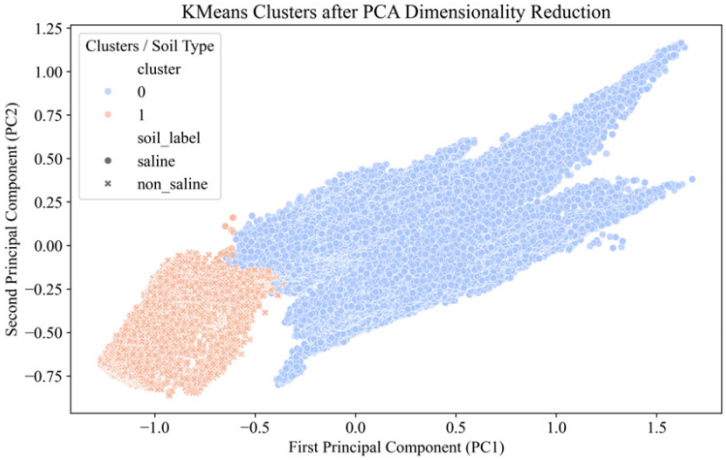
Нәтижелер және талқылау.

Ұсынылған әдіс топырақтың тұздануын болжау мен кеңістіктік жіктеуде жоғары тиімділікті көрсетті. Модель Sentinel-2 мультиспектрлік суреттері мен ERA5-Land климаттық деректерін біріктіріп, топырақтың спектралдық және гидрологиялық қасиеттерін кешенді түрде талдауға мүмкіндік берді. MSI сенсорының 10 метрлік ажыратымдылығы бар B2–B12 диапазондары өсімдік жамылғысы мен ылғалдылық көрсеткіштерін дәл сипаттады. Алынған SI1–SI5, NDVI, NDWI, NDSI және BI индекстері тұздану деңгейін анықтауда маңызды рөл атқарды. Алынған мәліметтерді терең талдау үшін K-Means кластерлеу әдісі қолданылып, топырақ үлгілері екі негізгі топқа — тұзданған (saline) және тұзданбаған (non_saline) кластерлерге бөлінді. 1-суретте K-Means алгоритмі арқылы анықталған кластерлер бойынша тұзданған және тұзданбаған топырақтардың таралуын көрсететін жылу картасы (heatmap) ұсынылған. Картада әрбір ұяшық белгілі бір кластерге жататын пиксельдердің санын бейнелейді. Нәтижелерге сәйкес, кластер 0 негізінен тұзданған топырақтарға, ал кластер 1 тұзданбаған топырақтарға сәйкес келеді. Кестеде көрсетілгендей, кластер 0 құрамында 2 405 794 тұзданған пиксель және 49 тұзданбаған пиксель бар, ал кластер 1 құрамында 40 552 тұзданбаған және 56 тұзданған пиксель кездеседі. Бұл нәтиже K-Means алгоритмінің тұздану дәрежесін ажыратуда жоғары дәлдікке ие екенін және екі кластердің нақты шекарамен бөлінетінін дәлелдейді.



***Сур.** 1.* K-Means кластерлеу нәтижесінде алынған тұзданған және тұзданбаған топырақтардың таралу картасы.

Көрсетілгендей, тұзданған және тұзданбаған топырақтар нақты екі кластерге бөлінеді, бұл олардың спектралдық және гидрологиялық айырмашылықтарын дәлелдейді. Келесі кезеңде PCA (Principal Component Analysis) әдісі қолданылып, алынған деректер екі басты компонент (PC1 және PC2) бойынша визуализацияланды, бұл Sentinel-2 MSI суреттері арқылы тұздану деңгейін бағалаудың тиімділігін көрсетті (Tashpolat & Reheman, 2025). 2-суретте PCA (Principal Component Analysis) әдісі қолданылып, алынған деректер екі басты компонент (PC1 және PC2) бойынша визуализацияланды. Бұл әдіс жоғары өлшемді деректердің құрылымын екі өлшемде көрсетуге және олардың ішкі байланыстарын түсінуге мүмкіндік береді. 2-суретте PCA нәтижелерінің негізінде K-Means кластерлеу алгоритмі арқылы бөлінген топырақ үлгілерінің таралуы көрсетілген. Ось X бірінші басты компонентті (PC1), ал ось Y екінші басты компонентті (PC2) білдіреді. Нүктелердің түсі кластер нөміріне (0 және 1), ал маркерлердің пішіні нақты топырақ түріне (Saline және Non_saline) сәйкес келеді.

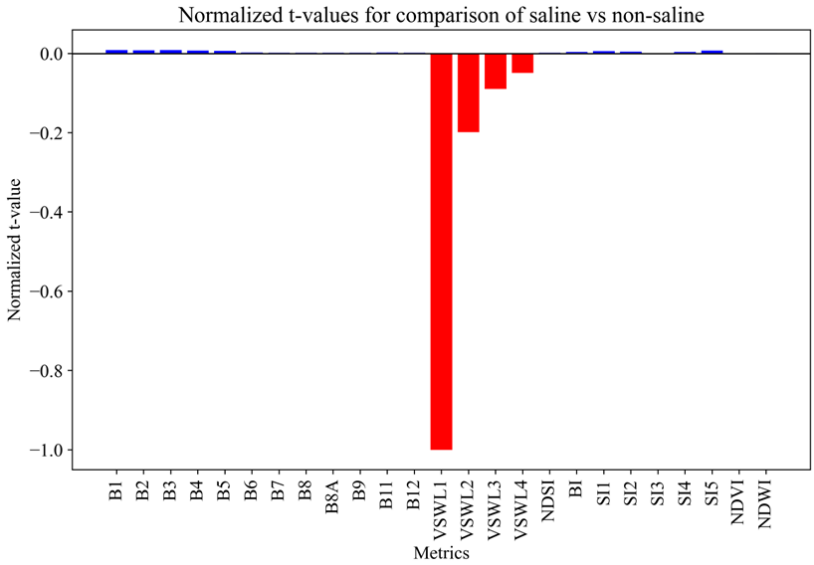


Сур. 2. PCA нәтижелері негізінде алынған K-Means кластерлерінің визуализациясы.

Графиктен байқағандай, тұзданбаған (Non_saline) топырақтар

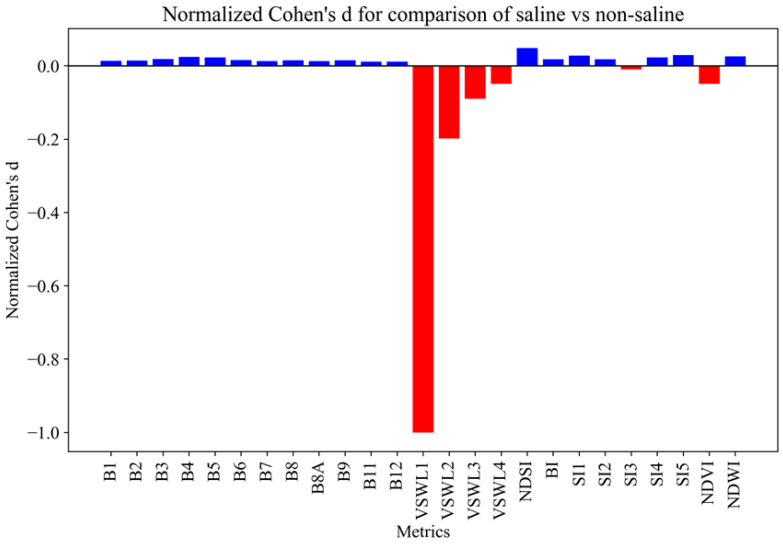
диаграмманың сол жағында, ал тұзданған (Saline) топырақтар оң жағында орналасқан, бұл олардың спектралдық және гидрологиялық айырмашылықтарының айқын екенін дәлелдейді. Сонымен қатар, тұзданған топырақ үлгілерінің таралуы анағұрлым кең, бұл олардың спектралдық көрсеткіштер мен ылғалдылық бойынша жоғары өзгергіштігін көрсетеді. Осылайша, PCA және K-Means әдістерін біріктіру топырақ типтерінің табиғи айырмашылықтарын тиімді анықтауға мүмкіндік береді.

3-суретте тұзданған (Saline) және тұзданбаған (Non_saline) топырақ үлгілерінің айырмашылығын көрсету үшін t-тест нәтижелерінің нормаланған мәндері берілген. Диаграмма екі топтың спектралдық және гидрологиялық параметрлері бойынша статистикалық айырмашылығын көрсетеді. Графиктен байқалғандай, B1–B12, SI1–SI5, BI, NDWI және NDSI индекстері оң t-мәндерге ие, яғни олардың мәні тұзданған топырақтарда жоғары. Ал VSWL1–VSWL4, NDVI және SI3 индекстері теріс мәнге ие, бұл олардың тұзданбаған топырақтарда жоғары екенін білдіреді. Бұл айырмашылықтар топырақтың физикалық және химиялық қасиеттерінің айқын ерекшеліктерін дәлелдейді.



Сур. 3. Тұзданған (Saline) және тұзданбаған (Non_saline) топырақ үлгілерінің салыстырмалы t-мәндері.

4-суретте тұзданған (Saline) және тұзданбаған (Non_saline) топырақ үлгілерінің арасындағы айырмашылықтарды сипаттайтын Cohen’s d мәндерінің нормаланған диаграммасы көрсетілген. Бұл көрсеткіш екі топ арасындағы айырмашылықтың шамасын анықтауға мүмкіндік береді. B1–B12, SI1–SI5, BI, NDWI және NDSI параметрлері оң мәндерге ие болып, тұзданған топырақтарда жоғары екенін көрсетеді, ал VSWL1–VSWL4, NDVI және SI3 теріс мәндерге ие, бұл олардың тұзданбаған топырақтарда жоғары екенін білдіреді. Осылайша, Cohen’s d әдісі топырақтардың физикалық және спектралдық қасиеттері арасындағы айқын айырмашылықтарды сандық тұрғыдан дәлелдейді.



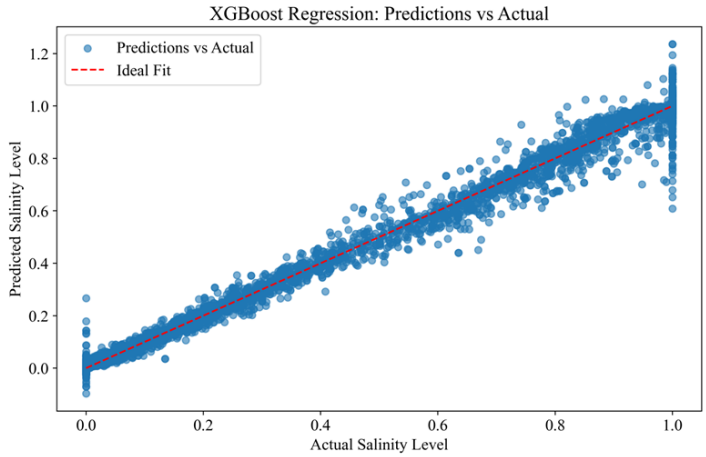
Сур. 4. Тұзданған (Saline) және тұзданбаған (Non_saline) топырақ үлгілерінің салыстырмалы Cohen’s d мәндері

4-суретте көрсетілгендей, тұзданған және тұзданбаған топырақтардың арасындағы айырмашылықтар айқын көрінеді. Көрсетілгендей, тұзданған топырақтарда спектралдық индекстердің мәндері жоғары, ал тұзданбаған топырақтарда ылғалдылық және өсімдік индексі (NDVI) жоғары. Бұл нәтижелер Cohen’s d әдісінің топырақтардың физикалық және спектралдық қасиеттеріндегі айырмашылықтарды тиімді анықтай алатынын дәлелдейді.

Алынған статистикалық талдау нәтижелері топырақ тұздылығының негізгі белгілерін анықтауға мүмкіндік берді. Келесі кезеңде осы белгілер XGBoost моделі арқылы тұздылық деңгейін болжау үшін пайдаланылды.

5-суретте XGBoost моделінің нақты (Actual) және болжамды (Predicted) тұздылық деңгейлері арасындағы тәуелділік көрсетілген. График нүктелер түрінде модельдің болжау нәтижелерін және қызыл пунктир сызықпен идеалды сәйкестік сызығын бейнелейді. Көрсетілгендей, нүктелердің басым бөлігі бұл сызыққа жақын орналасқан, бұл XGBoost моделінің жоғары дәлдікпен жұмыс істейтінін дәлелдейді. Дегенмен, төмен және жоғары диапазондарда аздаған ауытқулар байқалады, бұл модельдің шеткі мәндерде қателігі бар екенін көрсетеді. Жалпы алғанда, XGBoost моделі тұздылық деңгейін болжауда жоғары тиімділікті көрсетті.

XGBoost моделінің нақты (Actual) және болжамды (Predicted) тұздылық деңгейлері арасындағы тәуелділік көрсетілген. Бұл график модель болжаған мәндер мен нақты мәндердің өзара сәйкестігін нүктелер түрінде бейнелейді және қызыл пунктир сызық идеалды сәйкестікті білдіреді. Көрсетілгендей, нүктелердің басым бөлігі бұл сызыққа жақын орналасқан, бұл XGBoost моделінің жоғары дәлдікпен жұмыс істейтінін дәлелдейді.

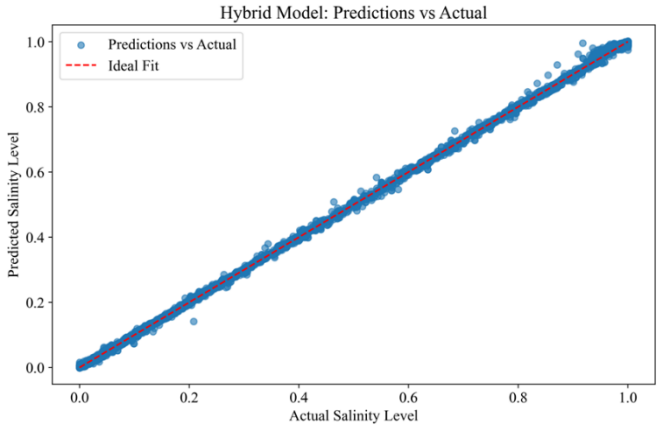


Сур. 5. XGBoost моделінің нақты және болжамды тұздылық деңгейлерін салыстыру диаграммасы.

Алайда төмен және жоғары диапазондарда аздаған ауытқулар байқалады, бұл модельдің шеткі мәндерде қателігі бар екенін көрсетеді. Орташа квадраттық қателік (MSE) мәні 0.0123 құрап, қанағаттанарлық нәтижені көрсетті, ал детерминация коэффициенті (R^2) болжау дәлдігінің жоғары екенін растады. Бұл нәтижелер XGBoost моделінің регрессиялық есептерді тиімді шешуге қабілетті екенін дәлелдейді. Дегенмен, модель PCA немесе логарифмдік түрлендіру сияқты қосымша өңдеу әдістерін пайдаланбағандықтан, оның жетілдіру әлеуеті сақталады. Жалпы алғанда, XGBoost әдісі маңызды белгілерді анықтауда жоғары тиімділік танытты және болжау дәлдігін одан әрі арттыру күрделірек гибридтік тәсілдерді қолдану арқылы мүмкін екенін көрсетті.

6-суретте гибридті модельдің нәтижелері көрсетілген. Бұл модель XGBoost алгоритмін логарифмдік түрлендіру, корреляциялық талдау және басты компоненттер әдісі (PCA) сияқты алдын ала өңдеу тәсілдерімен біріктіру арқылы жетілдірілді. Мұндай тәсіл базалық XGBoost моделінің шектеулерін жою және болжау дәлдігін арттыру мақсатында әзірленді. Графикте болжамды (Predicted) және нақты (Actual) тұздылық деңгейлері арасындағы тәуелділік бейнеленген. Көрсетілгендей, нүктелердің басым бөлігі идеалды сәйкестік сызығына (қызыл пунктир) өте жақын орналасқан, бұл гибридті модельдің жоғары дәлдікпен жұмыс істейтінін дәлелдейді. Алынған нәтижелер MSE мәнінің айтарлықтай төмендегенін және R^2 көрсеткішінің жақсарғанын көрсетті, бұл модельдің топырақ тұздылығын болжаудағы тиімділігін растайды

Көрсетілгендей, графикте болжау нәтижелері (Predicted) мен нақты тұздылық мәндері (Actual) арасындағы тәуелділік бейнеленген. Болжау нүктелері базалық XGBoost моделіне қарағанда идеалды сәйкестік сызығына (қызыл пунктир) айтарлықтай жақын орналасқан, бұл гибридті модельдің жоғары дәлдігін көрсетеді. Гибридті модель үшін орташа квадраттық қате (MSE) 0.0091 мәнін құрады, бұл XGBoost моделіне қарағанда шамамен 26%-ға төмен. .

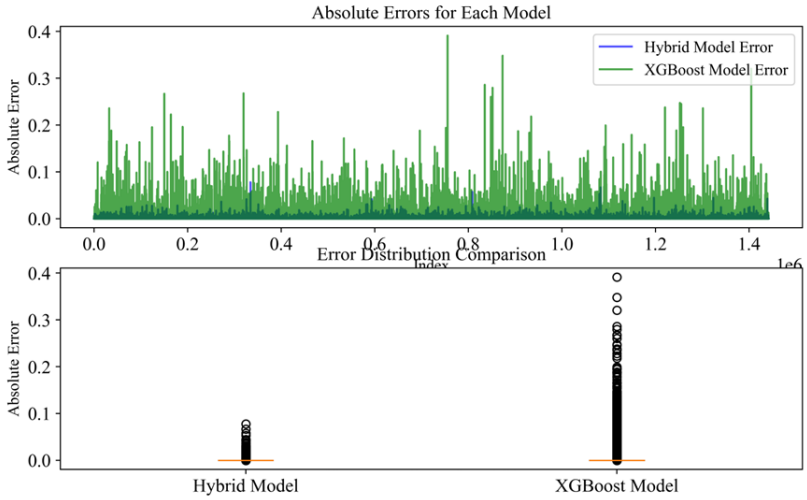


Сур. 6. Гибридті модельдің нақты және болжамды тұздылық деңгейлерін салыстыру диаграммасы.

Сонымен қатар, детерминация коэффициенті (R^2) жақсарып, модельдің деректер дисперсиясының басым бөлігін сәтті түсіндіретінін растады. Гибридті модельдің ерекшелігі – басты компоненттер әдісін (PCA) пайдалану, ол бастапқы белгілерді жаңа кеңістікке түрлендіріп, олардың толық өлшемділігін сақтауға мүмкіндік берді. Бұл тәсіл деректердің интерпретациясын жақсартты және модельге белгілер арасындағы тәуелділіктерді неғұрлым тиімді анықтауға мүмкіндік берді. Бұдан бөлек, логарифмдік түрлендіру шеткі мәндердің әсерін азайтып, болжау сапасын арттырды. Графиктен шеткі диапазондардағы ауытқулардың азайғаны және нүктелердің басым көпшілігінің идеалды сәйкестік сызығы бойында шоғырланғаны анық байқалады.

7-суретте гибридті және XGBoost модельдерінің абсолюттік қателіктерінің салыстырмалы графигі көрсетілген. Графиктен көрініп тұрғандай, гибридті модельде қателік мәндері біршама төмен және біркелкі таралған, ал XGBoost моделінде жоғары амплитудалы тербелістер мен шеткі мәндер жиірек кездеседі. Бұл гибридті тәсілдің болжам тұрақтылығын арттырып, модельдің жалпы өнімділігін жақсартқанын көрсетеді. Boxplot диаграммалары да гибридті модельде орташа және медианалық қателіктердің аз екенін айқын көрсетеді. Бұл нәтиже гибридті тәсілдің шулы және өзгермелі деректермен жұмыс істеуде анағұрлым тиімді екенін дәлелдейді.

Көрсетілгендей, гибридті және XGBoost модельдерінің қателіктерін салыстыру нәтижелері гибридті тәсілдің айқын артықшылығын көрсетті. Гибридті модельдің орташа абсолюттік және квадраттық қателік мәндері төмен болып, болжам дәлдігі мен тұрақтылығын арттырды. Бұл нәтижелер модельдің шулы және күрделі деректермен тиімді жұмыс істей алатынын дәлелдейді. Осылайша, жүргізілген талдау нәтижелері ұсынылған гибридті модельдің топырақтың тұздану деңгейін болжауда жоғары дәлдік пен тұрақтылыққа ие екенін көрсетті. Осылайша, жүргізілген талдау нәтижелері ұсынылған гибридті модельдің топырақтың тұздану деңгейін болжауда жоғары дәлдік пен тұрақтылыққа ие екенін көрсетті.



Сур. 7. Гибридті және XGBoost модельдерінің абсолюттік қателіктерінің салыстырмалы графигі.

Гибридті тәсіл спектралдық және гидрологиялық деректерді біріктіре отырып, дәстүрлі XGBoost моделіне қарағанда болжам сапасын айтарлықтай жақсартты. Нәтижелер MAE және RMSE қателерінің төмен мәндерімен, ал R^2 көрсеткішінің жоғары болуымен расталды. Сонымен қатар, PCA және корреляциялық талдау сияқты алдын ала өңдеу әдістерін қолдану модельдің бейімделу қабілетін арттырып, тұзданған және тұзданбаған топырақтар арасындағы айырмашылықты неғұрлым дәл сипаттауға мүмкіндік берді. Жалпы алғанда, ұсынылған модель кеңістіктік жіктеу және тұздылықты сандық бағалау міндеттерін шешуде тиімді және перспективалы құрал болып табылады.

Қорытынды

Зерттеу барысында Sentinel-2 спутниктік мультиспектрлік деректері мен ERA5-Land гидрологиялық параметрлерін біріктіру арқылы топырақтың тұздылығын кешенді бағалау жүргізілді (Zhao et al., 2024; Hersbach et al., 2023). Бұл тәсіл тұзданған (Saline) және тұзданбаған (Non_saline) топырақтардың айырмашылықтарын сипаттайтын негізгі заңдылықтарды анықтауға мүмкіндік берді (Jiang et al., 2025). t-тест пен Cohen's d әсер шамасы арқылы жүргізілген статистикалық талдау спектралдық индекстердің (NDSI, SI4, BI) және топырақ ылғалдылығының параметрлерінің (VSWL1–VSWL4) топырақ тұздылығын анықтайтын негізгі факторлар екенін көрсетті (Cohen, 1988). Тұзданған топырақтарда спектралдық индекстердің жоғары мәндері және ылғалдың төмен деңгейі байқалса, тұзданбаған топырақтарда керісінше NDVI мәні жоғары және ылғалдылығы жақсы болды (Cui et al., 2025).

Басты компоненттер әдісін (PCA) қолдану деректердің өзгергіштігін сипаттайтын екі негізгі бағытты анықтауға мүмкіндік берді (Jolliffe & Cadima, 2016; Tashpolat & Reheman, 2025): гидрологиялық және спектралдық. Бұл тәсіл K-Means кластерлеу әдісі арқылы екі топты (тұзданған және тұзданбаған)

98%-дан астам дәлдікпен ажыратуға мүмкіндік берді (Becker et al., 2025). Келесі кезеңде XGBoost алгоритміне негізделген гибридті модель әзірленіп, логарифмдік түрлендіру, корреляциялық талдау және PCA сияқты алдын ала өңдеу әдістері қолданылды (Chen & Guestrin, 2016; Wang & Xu, 2023). Бұл модель базалық нұсқамен салыстырғанда орташа квадраттық қателікті (MSE) 26%-ға төмендетіп, тұздылықтың шеткі мәндерін болжаудағы дәлдігін арттырды.

Гидрологиялық параметрлер мен спектралдық көрсеткіштерді біріктіру модельдің дәлдігін және интерпретациясын жақсартты (Gao & Xu, 2025), ал тұздылық пен ылғалдылық арасындағы күрделі тәуелділіктерді анықтауға мүмкіндік берді. Сонымен қатар, модельге енгізілген бұлдыр логика (fuzzy logic) тұздылық деңгейлерінің шекараларын икемді түрде анықтауға жағдай жасап, оны аймақтық айырмашылықтарға бейімдеді (Li et al., 2022).

Осылайша, ұсынылған гибридті тәсіл топырақ тұздылығын бақылау мен болжауда тиімділігін дәлелдеді. Ол жоғары дәлдікпен автоматты классификация жүргізіп, шулы деректермен де тұрақты жұмыс істеуге мүмкіндік береді. Ұсынылған әдіс экологиялық мониторинг, жер ресурстарын басқару және агроэкологиялық зерттеулер салаларында қолдануға болатын әмбебап құрал ретінде ерекшеленеді (Luo et al., 2025).

REFERENCES

Becker, S.J., Maloney, M.C., Griffin, A.W., Lasko, K. & Sussman, H.S. (2025). Bare ground classification using a spectral index ensemble and machine learning models optimized across 12 international study sites // Geocarto International. — 2025. — Vol. 40. — No. 1. — Article 2465452. DOI: 10.1080/10106049.2023.2465452.

Bo, Q., Lv, P., Wang, Z., Wang, Q., & Li, Z. (2024). Prediction of the post-mining land use based on random forest and DBSCAN // PLoS ONE. — 2024. — Vol. 19. — No. 1. — Article e0287079. DOI: 10.1371/journal.pone.0287079.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2016. — Pp. 785–794. DOI: 10.1145/2939672.2939785.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). — Lawrence Erlbaum Associates.

Copernicus Open Access Hub. (2025). Sentinel-2 User Guide // European Space Agency (ESA). — 2025. — URL: https://sentiwiki.esa.int/sentinel-2.

Cui, G., Liu, Y., Li, X., Wang, S., Qu, X., Wang, L., & Zhang, W. (2025). Impacts of groundwater storage variability on soil salinization in a semi-arid agricultural plain // Geoderma. — 2025. — Vol. 454. — Article 117162.

Gao, L., & Xu, E. (2025). Mapping cropland soil salinity using multi-cycle classification to mitigate retrieval errors // Computers and Electronics in Agriculture. — 2025. — Vol. 231. — Article 110055. DOI: 10.1016/j.compag.2024.110055.

Hersbach, H., Bell, B., Berrisford, P., et al. (2023). ERA5-Land hourly data from 1950 to present // Copernicus Climate Change Service (C3S) Climate Data Store (CDS). — 2023. — DOI: 10.24381/cds.e2161bac.

Jiang, Z., Ding, J., Li, Z., & Liu, J. (2025). Present knowledge and future challenges in remote sensing for soil salinization monitoring: A review of bibliometric analysis // International Journal of Remote Sensing. — 2025. — Vol. 46. — No. 1. — Pp. 247–272. DOI: 10.1080/01431161.2024.2412804.

Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: A review and recent developments // Philosophical Transactions of the Royal Society A. — 2016. — Vol. 374. — No. 2065. — Article 20150202. DOI: 10.1098/rsta.2015.0202.

Li, H., Zhang, J., & Ren, S. (2022). Fuzzy logic approach for soil salinity classification using remote sensing data // Catena. — 2022. — Vol. 210. — Article 105933. DOI: 10.1016/j.catena.2021.105933.

Luo, Z., Deng, M., Tang, M., Liu, R., Feng, S., Zhang, C., & Zheng, Z. (2025). Estimating soil profile salinity

under vegetation cover based on UAV multi-source remote sensing // Scientific Reports. — 2025. — Vol. 15. — No. 1. — Article 2713. DOI: 10.1038/s41598-024-40157-9.

Tashpolat, N., & Reheman, A. (2025). Monitoring of soil salinity in the Weiku Oasis based on feature space models derived from Sentinel-2 MSI images // Land. — 2025. — Vol. 14. — No. 2. — Article 251. DOI: 10.3390/land14020251.

Wang, S., & Xu, D. (2023). Hybrid models integrating PCA and XGBoost for soil salinity prediction // Environmental Science and Pollution Research. — 2023. — Vol. 30. — No. 6. — Pp. 15012–15026.

Zhao, F., Liu, Y., & Zhang, C. (2024). Multi-source data fusion for soil salinity assessment using ERA5 and Sentinel-2 datasets // Remote Sensing Applications: Society and Environment. — 2024. — Vol. 34. — Article 101102.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 114–135
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.007

UDC 004.8:336.76

FORECASTING THE FINANCIAL CONDITION OF ENTERPRISES USING FUZZY LOGIC AND CLUSTER ANALYSIS METHODS

A. Bykov[1*], A. Yemberdiyeva[1], N. Daurenbayeva[1], A. Nurlanuly[2]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Civil Aviation Academy.

Artem Bykov — c.t.s., Associate Professor, Head of Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
E-mail: a.bykov@iitu.edu.kz, https://orcid.org/0000-0002-9563-5185;
Aknur Yemberdiyeva — master, senior-lecturer, Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0009-0005-5078-2412;
Nurkamilya Daurenbayeva — doctoral student, Master's degree, assistant professor, International University of Information Technology, Almaty, Kazakhstan
https://orcid.org/0000-0003-0341-4017;
Almas Nurlanuly — doctoral student, Master of Technical sciences, Civil Aviation Academy, senior lecturer of the Department of Aviation Engineering and Technology, Almaty, Kazakhstan
https://orcid.org/0000-0002-0364-0455.

Abstract. The analysis of indicators of complex economic systems operating under significant uncertainty and incomplete statistical information requires the use of specialized methodological tools. In such cases, traditional financial analytics often fails to provide accurate results, especially when non-financial data must be considered. One of the most effective approaches is the application of fuzzy logic theory, which allows recognition and forecasting of financial conditions even with contradictory or insufficient data. The proposed methodology consists of two sequential stages: the development of membership functions and the clustering of indicators within the space of financial conditions. As a result, enterprises can be classified into two main groups: normal operation and pre-bankruptcy state. Experimental research has shown that the reliability of this classification is comparable to neural network methods. At the same time, the approach maintains an important advantage — expert involvement in constructing membership functions, which ensures satisfactory recognition out-

comes under uncertainty and the influence of various external factors.

Keywords: financial condition, bankruptcy forecasting, fuzzy logic, membership function, clustering, equivalence classes, similarity relation

For citation: A. Bykov, A. Yemberdiyeva, N. Daurenbayeva, **A.** Nurlanuly. Forecasting the financial condition of enterprises using fuzzy logic and cluster analysis methods. 2025. Vol. 6. No. 24. Pp. 114–135. (In Kaz.). https://doi.org/10.54309/IJICT.2025.24.4.007.

Conflict of interest: The authors declare that there is no conflict of interest.

АНЫҚ ЕМЕС ЛОГИКА ЖӘНЕ КЛАСТЕРЛІК ТАЛДАУ ӘДІСТЕРІН ҚОЛДАНА ОТЫРЫП, КӘСІПОРЫНДАРДЫҢ ҚАРЖЫЛЫҚ ЖАҒДАЙЫН БОЛЖАУ

А.Быков[1*], А.Ембердіева[1], Н.Дауренбаева[1], А.Нұрланұлы[2]
Халықаралық Ақпараттық Технологиялар Университеті, Алматы,Қазақстан;
[2]Азаматтық авиация академиясы, Алматы,Қазақстан.
E-mail: a.bykov@iitu.edu.kz

Артем Быков — т.ғ.к., доцент, кафедра меңгерушісі, Компьютерлік инженерия кафедрасы, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
E-mail:a.bykov@iitu.edu.kz, https://orcid.org/0000-0002-9563-5185;
Ақнұр Ембердіева — магистр, сениор-лектор, Компьютерлік инженерия кафедрасы, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
https://orcid.org/0009-0005-5078-2412;
Нұркамиля Дауренбаева — докторант, магистр, ассистент-профессор, Компьютерлік инженерия кафедрасы, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-0341-4017;
Алмас Нұрланұлы — докторант, техника ғылымдарының магистрі, Азаматтық авиация академиясы, Авиациялық техника және технологиялар кафедрасының сениор лекторы, Алматы, Қазақстан
https://orcid.org/0000-0002-0364-0455.

Аннотация. Маңызды белгісіздік жағдайында жұмыс істейтін күрделі құрылымдалған экономикалық жүйелердің көрсеткіштерін талдау, қандай да бір толыққанды статистика болмаған кезде немесе зерттелетін көрсеткіштердің қатарына қаржылық емес деректерді қосу қажет, қаржылық жағдайды талдаудың арнайы әдіснамалық құралдарын қолдануды талап етеді. Шешілетін мәселелерге сәйкестігін дәлелдеген осындай құралдардың бірі-бұлыңғыр логика

теориясының математикалық аппараты. Кез-келген экономикалық жүйелерді аналитикалық қамтамасыз етудің негізі қаржылық аналитика әдіснамасы болып табылады, әдетте әр түрлі белгісіздік жағдайында қаржылық жағдайларды тану және болжау құралдарын қамтиды. Мақалада кәсіпорынның қаржылық жағдайын талдау әдістемесі келтірілген, ол дәйекті түрде орындалатын екі процедураны қамтиды: 1) тиесілілік функцияларын қалыптастыру және 2) қаржылық жағдай индикаторларының кеңістігінде кластерлеу. Сол үлгідегі көп қабатты перцептронның нейрондық желілік архитектурасымен салыстыру көрсеткендей, ұсынылған әдіс MLP-дегі 91,2 % - ға қарсы 89,4 % дәлдікке қол жеткізді, сонымен бірге интерпретация мен толық емес деректерге төзімділікті сақтап қалды. Әдістеменің эксперименттік зерттеулері көрсеткендей, екі кластердің біріне (қалыпты және банкроттыққа дейінгі күй) жатқызудың сенімділігі нейрондық желі әдістерімен салыстыруға болады. Сонымен қатар, тиістілік функцияларын құрастыру үшін сарапшыларды тартуға байланысты Әдістеменің артықшылықтары сақталады – бұл қарама-қайшы, толық емес деректер мен белгісіздік факторларының әртүрлі факторлары жағдайында танудың қанағаттанарлық нәтижесін қамтамасыз ету мүмкіндігі.

Түйін сөздер: қаржылық жағдай, банкроттықты болжау, бұлыңғыр логика, тиістілік функциясы, кластерлеу , эквиваленттік сыныптар, ұқсастық қатынасы

Дәйексөздер үшін: А. Быков, А. Ембердіева, Н. Дауренбаева, А. Нұрланұлы. Анық емес логика және кластерлік талдау әдістерін қолдана отырып, кәсіпорындардың қаржылық жағдайын болжау // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 114–135 бет. (Қаз). https://doi.org/10.54309/IJICT.2025.24.4.007.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ПРОГНОЗИРОВАНИЕ ФИНАНСОВОГО СОСТОЯНИЯ ПРЕДПРИЯТИЙ С ПРИМЕНЕНИЕМ МЕТОДОВ НЕЧЕТКОЙ ЛОГИКИ И КЛАСТЕРНОГО АНАЛИЗА

А. Быков[1*], А. Ембердіева[1], Н. Дауренбаева[1], А. Нұрланұлы[2]

[1] Международный университет информационных технологий, Алматы, Казахстан;

[2]Академия гражданской авиаций (АГА), Алматы, Казахстан.
E-mail: a.bykov@iitu.edu.kz

Артем Быков — к.т.н., доцент, заведующий кафедрой «Компьютерная инженерия», Международный Университет Информационных технологии, Алматы, Казахстан
E-mail: a.bykov@iitu.edu.kz, https://orcid.org/0000-0002-9563-5185;

Акнур Ембердиева — магистр, сениор-лектор, кафедра «Компьютерная инженерия», Международный Университет Информационных технологии, Алматы, Казахстан
https://orcid.org/0009-0005-5078-2412;
Нұркамиля Дауренбаева — докторант, магистр, ассистент профессор, Международный университет информационных технологий, кафедра «Компьютерная инженерия», Алматы, Казахстан
https://orcid.org/0000-0003-0341-4017;
Алмас Нұрланұлы — докторант, магистр технических наук, Академия гражданской авиаций (АГА), сениор лектор кафедры «Авиационная техника и технология», Алматы, Қазақстан
https://orcid.org/0000-0002-0364-0455.

Аннотация. Анализ показателей сложно-структурируемых экономических систем, функционирующих в условиях существенной неопределенности, когда отсутствует сколько-нибудь полноценная статистика, или в число исследуемых показателей необходимо включать нефинансовые данные, требует применения особого методологического инструментария анализа финансового состояния. Одним из таких инструментариев, доказавшим свою адекватность решаемым задачам является математический аппарат теории нечеткой логики. Основу аналитического обеспечения любой экономических систем составляет методология финансовой аналитики, как правило включающая инструментарий распознавания и прогнозирования финансовых состояний в условиях различной степени неопределенности. В статье представлена методика анализа финансового состояния предприятия, включающая две последовательно выполняемый процедуры: 1) формирования функций принадлежности и 2) кластеризации в пространстве индикаторов финансовых состояний. Как показали экспериментальные исследования методики, достоверность отнесения к одному из двух кластеров (нормальное и предбанкротного состояния) сравнима с нейросетевыми методами. В тоже время сохраняются преимущества методики, обусловленное привлечением экспертов для составления функций принадлежности – это возможность обеспечения удовлетворительного результата распознавания в условиях противоречивых, неполных данных и различного рода факторах неопределенности.

Ключевые слова: финансовое состояние, прогнозирование банкротства, нечеткая логика, функция принадлежности, кластеризация, классы эквивалентности, отношение подобия

Для цитирования: А. Быков, А. Ембердіева, Н. Дауренбаева, А. Нұрланұлы. прогнозирование финансового состояния предприятий с применением методов нечеткой логики и кластерного анализа//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24.

Стр. 114–135. (На каз.). https://doi.org/10.54309/IJICT.2025.24.4.007.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе

Кәсіпорынның қаржылық жағдайы ресурстармен (материалдық, адами) қамтамасыз етілуімен, жаңа технологияларды (физикалық және заңды) иеленуімен, экономикалық қызметтің басқа субъектілерімен экономикалық қатынастарымен, төлем қабілеттілігімен және тұрақтылығымен сипатталады ().

Экономикалық ғылымда бұл ұғымды біркелкі анықтайтын ғылыми негізделген тәсіл әлі қалыптаспаған (). Көптеген зерттеулерде кәсіпорынның қаржылық жағдайы оның өз қызметін қаржыландыру қабілеттілігінің деңгейі ретінде сипатталады. Бұл көрсеткіш ұйымның қаржылық ресурстармен қамтамасыз етілуін, оларды тиімді орналастыруын, қолдану қарқындылығы мен тиімділігін, сондай-ақ бағалы қағаздар нарығындағы тұрақтылығын қамтиды. Ашық акционерлік қоғамдар үшін бұл факторлар акциялардың бағасына, инвесторлардың сеніміне және капиталдың тартымдылығына тікелей әсер етеді.

Кәсіпорынның қаржылық жағдайын талдау оның қызметін тиімді басқару үшін аналитикалық қамтамасыз етуді қалыптастырудың негізгі аспектісі болып табылады (). Сонымен қатар, кез-келген экономикалық жүйенің аналитикалық қамтамасыз етілуінің негізі қаржылық аналитика әдістемесіне сүйенеді. Бұл әдістеме әртүрлі белгісіздік жағдайларында жүйенің қаржылық жағдайын бағалау, тану және болжау құралдарын қамтиды ().

Қазіргі таңда ықтималдық-статистикалық тәсілмен қатар, бұлыңғыр логика негізіндегі әдістер кеңінен қолданылуда (Бернстайн, 2015). Бұл тәсіл белгісіздік жағдайында ақпаратты өңдеуге және болжам жасаудың дәлірек әдістерін қалыптастыруға мүмкіндік береді. Белгілі артықшылықтарының арқасында, бұлыңғыр логика азық-түлік қауіпсіздігін басқару, қаржылық жүйелерді талдау, өндірістік процестерді оңтайландыру сияқты әртүрлі күрделі технологиялық процестерде тиімді қолданылып келеді. Бұл тәсіл әсіресе «кіру-шығу» тәуелділігі нақты анықталмаған, күрделі жүйелерде ерекше пайдалы болып табылады (Carvalho және т.б., 2009; Mukhanov және т.б., 2021; Mukhanov және т.б. 2024).

Кәсіпорынның қаржылық жағдайын дәл және объективті бағалау бизнесті басқарудың маңызды элементі болып табылады. Қаржылық жағдайдың тұрақтылығын немесе дағдарыстық күйін анықтау үшін компанияның балансын, қаржылық есептілігін, ақша ағындарын және негізгі экономикалық көрсеткіштерін талдау қажет.

Қаржылық жағдайды бағалау бірнеше негізгі аспектілерді қамтиды:

• Өтімділік – компанияның қысқа мерзімді міндеттемелерін өтеу қабілеті

• Төлем қабілеттілігі – ұйымның барлық қарыздық міндеттемелерін уақтылы орындау мүмкіндігі.

• Рентабельділік – компанияның кіріс алу тиімділігі мен оның активтерін пайдалану деңгейі.

• Қаржылық тұрақтылық – ұзақ мерзімді даму стратегиясының жүзеге асырылу мүмкіндігі.

Осы факторларды ескеру арқылы, кәсіпорынның қаржылық жағдайын тереңірек зерттеу мүмкін болады, бұл өз кезегінде банкроттықтың алдын алу және қаржылық тұрақтылықты сақтау үшін қажет.

Дәстүрлі қаржылық талдау әдістері белгілі бір шектеулерге ие, себебі олар нақты және толық деректердің болуын талап етеді. Алайда, көптеген жағдайларда кәсіпорынның қаржылық жағдайын сипаттайтын деректер толық емес немесе қарама-қайшы болуы мүмкін. Мұндай белгісіздік жағдайында бұлыңғыр логикаға негізделген әдістер тиімді шешім ұсына алады.

Бұлыңғыр логиканың негізгі артықшылықтары:

• Белгісіз және толық емес деректермен жұмыс істеу мүмкіндігі.

• Кәсіпорынның қаржылық жағдайын дәлірек модельдеу.

• Стандартты әдістермен қиын анықталатын жасырын заңдылықтарды табу.

• Сараптамалық бағалауларды талдау және оларды шешім қабылдауда пайдалану.

Бұлыңғыр логика қатаң математикалық шектеулерге тәуелді емес, сондықтан ол динамикалық өзгерістерге икемді және сараптамалық тәсілдерді ескере отырып, дәлірек болжам жасауға мүмкіндік береді.

Зерттеудің мақсаты – Ягер процедурасын өзгерту және кейіннен бұлыңғыр кластерлеу негізінде тиесілік дәрежесін қалыптастыру әдістемесі негізінде кәсіпорындардың қаржылық жағдайы кеңістігіндегі “норма/банкроттық” дихотомиясындағы кластерлеудің дұрыстығын арттыру. Атап айтқанда, анық емес логиканың математикалық әдістермен және реттеуші болжаудың ұйымдастырылған процедурасының белгілі бір әдісімен бизнестің қаржылық жағдайын талдаудың нақты мәселелерін шешуде ықтималдық–статистикалық және анық емес мүмкін тәсілдер арасында байланыс орнатуға болады, мысалы, банкроттыққа дейінгі жағдайларды тану және болжау.

Бұл тәсіл экономикалық жүйелердің белгісіздік жағдайында жұмыс істеу ерекшеліктерін ескеруге мүмкіндік береді. Ягер процедурасын өзгерту кәсіпорындардың қаржылық көрсеткіштерінің күрделі құрылымын дәлірек бейнелеуге және мәліметтерді өңдеудің икемді моделін құруға жағдай жасайды. Бұлыңғыр кластерлеудің көмегімен қаржылық көрсеткіштердің вариациясын есепке алып, әрбір кәсіпорынның банкроттық қаупіне тиесілілік дәрежесін жоғары дәлдікпен анықтауға болады.

Сонымен қатар, зерттеуде кәсіпорындардың қаржылық жағдайын талдаудың инновациялық әдісі ұсынылады, ол дәстүрлі қаржылық коэффициенттер мен бұлыңғыр логиканың үйлесімдігін қамтамасыз етеді. Бұл тәсіл банкроттыққа дейінгі белгілерді алдын ала анықтауға және бизнес

субъектілерінің қаржылық орнықтылығын болжауға мүмкіндік береді. Бұлыңғыр жиындар теориясын қолдану арқылы зерттеу әртүрлі деңгейдегі экономикалық белгісіздіктерді ескеруге жағдай жасайды, бұл әсіресе тұрақсыз нарықтық ортада шешім қабылдау үдерісін оңтайландырады.

Әдістеменің тиімділігі эксперименттік зерттеулермен расталған, олардың нәтижелері көрсеткендей, бұлыңғыр кластерлеу негізінде кәсіпорындарды қаржылық тұрақтылық деңгейлері бойынша бөлу дәстүрлі статистикалық әдістермен салыстырғанда жоғары дәлдікке ие. Бұл зерттеу кәсіпорындардың қаржылық жағдайын бағалау және болжау саласында қолдануға болатын перспективалы аналитикалық құрал ұсынады.

Материалдар мен әдістер

Тиістілік дәрежесін алдын ала анықтау рәсімі. Егер кәсіпорынның қаржылық жағдайын талдау міндеті алдын-ала сипатталған немесе есептелген күй кластарын тану мағынасында скоринг міндеті ретінде ұсынылса, онда бұлыңғыр логика сенімсіз және әлсіз ресімделген мәліметтермен есептерді тиімді шешуге мүмкіндік береді, сонымен қатар жасанды интеллект пен дәстүрлі әдістер арасындағы аралық болып табылады (Helfert және т.б., 2015). Бұлыңғыр жиындар теориясының негізгі идеялары көпмағыналы логикада бұлыңғыр жиындар теориясы пайда болғанға дейін енгізілген ережелермен айтарлықтай сәйкес келеді (Muhd және т.б., 2013). Алайда, соңғысымен салыстырғанда бұлыңғыр жиындар теориясы практикада салыстырмалы түрде жоғары таралуымен сипатталады. Бұл түсініксіз математикада компьютерлік жүйелермен, диалогтық жүйелерді дамытумен, жасанды интеллект идеяларын енгізумен байланысты практикалық қосымшаларға бірден курс қабылданғандығымен түсіндіріледі. Бұл жерде Л.Заде бұлыңғыр жиындар бойынша зерттеулердің басынан бастап осы зерттеулер мен лингвистикалық айнымалы тұжырымдаманы біртұтас тұтастыққа біріктірген жағдай да маңызды рөл атқарды.

Бұлыңғыр логика теориясын қолдану кәсіпорындардың қаржылық жағдайын бағалауда сенімсіз немесе толық емес деректерді тиімді өңдеуге мүмкіндік береді. Бұл әсіресе экономикалық дағдарыстар, қаржы нарығындағы тұрақсыздық және макроэкономикалық өзгерістер жағдайында маңызды рөл атқарады. Дәстүрлі әдістер көбінесе қатал шектеулерге сүйенсе, бұлыңғыр логика деректердегі анықсыздықты және субъективті сараптамалық бағаларды ескере отырып, шешім қабылдауға мүмкіндік береді.

Кәсіпорындардың қаржылық тұрақтылығын болжау мен банкроттық тәуекелін бағалау міндеттері негізінен тиесілілік функцияларын құруды және мәліметтерді кластерлеуді қамтиды. Бұл тәсіл қаржылық көрсеткіштерді динамикалық талдауға мүмкіндік береді және компанияның даму үрдістерін неғұрлым нақты болжауға жағдай жасайды. Бұлыңғыр жиындар теориясының математикалық аппараты арқылы кәсіпорынның қаржылық жағдайы, оның тұрақтылық деңгейі, сондай-ақ банкроттық қаупі бар ма, жоқ па деген сұраққа жауап беруге болады.

Сонымен қатар, бұлыңғыр логика нейрондық желілер, статистикалық модельдер және машиналық оқыту әдістерімен біріктірілуі мүмкін. Мұндай гибридті әдістер шешім қабылдау процесінің тиімділігін арттырып, кәсіпорындардың қаржылық жағдайын тереңірек талдауға мүмкіндік береді. Бұлыңғыр логика деректерді алдын-ала стандартталған критерийлерге сәйкестендірмей-ақ өңдеуге мүмкіндік беретіндіктен, ол сарапшылардың пікірін, жинақталған статистиканы, сондай-ақ макроэкономикалық факторларды ескеруге жағдай жасайды.

Тану және болжау міндеттерінің көпшілігінің кезеңдерінің тізіміне зерттелетін экономикалық жүйелердің қаржылық жағдайларын жатқызу стандарттарын қалыптастыру үшін тиесілілік функциялары мен кластерлеу процедураларын қалыптастыру кезеңдері кіреді. Бұл әдіс компаниялардың қаржылық орнықтылығын болжау кезінде бірқатар дәстүрлі әдістердің кемшіліктерін жояды, себебі ол айнымалы мәндердің кең диапазонын қамтиды және әртүрлі сценарийлер бойынша қаржылық тұрақтылықты модельдеуге мүмкіндік береді. Бұлыңғыр логиканы қолдану кәсіпорындардың қаржылық тәуекелдерін жоғары дәлдікпен болжауға, сондай-ақ шешім қабылдаушыларға қосымша құралдар ұсынуға мүмкіндік береді.

Осылайша, бұлыңғыр логика негізінде ұсынылған әдістеме қаржылық жағдайды бағалау мен болжаудың жаңа мүмкіндіктерін ашады, ол қарама-қайшы немесе толық емес ақпаратты өңдеуге қабілетті және экономикалық жүйелердің күрделілігіне бейімделе алады. Бұл зерттеу бұлыңғыр логиканың кәсіпорындарды қаржылық диагностика мен болжау процесінде қолданылу аясын кеңейтуге негіз бола алады (Altman E., Hotchkiss E. (2006). , Ефанова Н.В., Ващенко В.Р. 2019).

Процедураны әзірлеу барысында біз жұптарды қамтитын X кеңістігінде бұлыңғыр a жиынтығының келесі классикалық анықтамасын ұстанамыз $\{(x, m_A(x))\}$, мұндағы $m_A(x)$ □ [0,1] - «x элементінің A жиынына жату дәрежесі A» (Mukhanov және т.б., 2023). Әлбетте, егер $m_A(x)$ функциясы тек 0 немесе 1 мәндерін қабылдау қасиетіне ие болса, онда бұлыңғыр жиынның берілген анықтамасы жиынның классикалық (канторлық) анықтамасымен сәйкес келеді, оған сәйкес не x O A , не x A орын алады.Соңғы талапты келесі түрдегі тиістілік функциясы арқылы да жазуға болады: m_A: $X \rightarrow \{0,1\}$. Бұл ережелер бұлыңғыр жиынтықтар теориясын әдеттегі жиынтық теорияны жалпылау ретінде қарастыруға болатындығын, сондықтан соңғысын ерекше жағдай ретінде қамтитындығын растайды. Өздеріңіз білетіндей, а жиынтығы осы қасиетке байланысты (мысалы, меншікті капитал мөлшері, қарыз және т.б.) белгілі бір қасиетке ие және X-тен бөлінетін элементтердің жиынтығы ретінде анықталады (Mukhanov және т.б., 2023). Бұлыңғыр жиынның әдеттегіден айырмашылығы, элементтің берілген қасиетке ие элементтер жиынтығына жататындығынан басқа, x элементтің $m_A(x)$ тиесілілік дәрежесі ретінде осы қасиетке ие болуының сандық бағасы да жазылады (Mukhanov және т.б., 2024).

Осы қасиеттен мүлдем айырылған x OX элементтері үшін $m_A(x)=0$ деп санайды және мұндай элементтер массивке кірмейді.

N элементтерден тұратын ақырлы бұлыңғыр жиын ,

$$A=\left\{\left(x_1,m_A\left(x_1\right)\right),\left(x_2,m_A\left(x_2\right)\right),...,\left(x_n,m_A\left(x_n\right)\right)\right\},\ \ x_i\in X \tag{1}$$

Ол сондай-ақ түрінде ұсынылады

$$A=m_1\big/_{x_1}+m_2\big/_{x_2}+,...,m_n\big/_{x_n},\ \ x_i\in X \tag{2}$$

Немесе

$$A=\sum_{i=1}^{n}m_i\big/_{x_i}=\bigcup_{i=1}^{n}m_i\big/_{x_i},\ \ x_i\in X. \tag{3}$$

Соңғы жаба (2) және (3) біріктіру ретінде түсіндірілуі керек екенін көрсетеді (арифметикалық жиынтық емес) және Σ-ды $\int$ -ға ауыстыру арқылы келесі жазбаны аламыз:

$$A=\int m_A\left(x\right)/x, \tag{4}$$

Мұндағы $m_A(x)$ – X кеңістігіндегі анықталған тиістілік функциясы

Төменде келтірілген процедура бұлыңғыр жиынның деңгей жиындарына ыдырауына негізделген. Тәсіл әмбебап жиын элементтерінің $m_A(x_i)$ мүшелігінің белгілі дәрежелері арасында байланыс орнататын формулаларды қолданады $X=\{x_1, x_2,... ,x_n\}$ бұлыңғыр жиынға А және қаржылық көрсеткіштерді бақылау және белгілі бір жолмен ұйымдастырылған шешімдер қабылдау кезінде осы $P(x_i)$ таңдау ықтималдығы. Бұл жағдайда Aa_1, Aa_2, …,Aa_i түрінің деңгейлік жиындары енгізіледі:

$$Aa=\left\{x\,|\,m_A\left(x\right)\geq a\right\}, \tag{5}$$

Мұндағы мүшелік дәрежесі А -дан үлкен немесе а – ға тең. Әрбір бұлыңғыр жиынды деңгей жиындарына бөлуге болады және сол жиындардың бірігуі ретінде ұсынылуы мүмкін, А сонымен қатар олар үшін $a_j=m(x_j)$ шартын анықтаңыз.

Қауымдастықты шектемей x_i элементтері мүшелік дәрежесінің өсу ретімен нөмірленеді деп санауға болады. Содан кейін Aa_i жиынтықтары келесі түрге ие болады:

$$Aa_1=X=\left\{x_1,x_2,...,x_n\right\},\ \ Aa_2=\left\{x_2,x_3,...,x_n\right\},\ \ Aa_n=\left\{x_n\right\}. \tag{6}$$

Қаржылық көрсеткіштерге бақылау жүргізу және олардың жай-күй класы туралы шешім қабылдау кезінде, жиынтықтан $[a_1,.a_2,\ldots,a_n]$ a_j элементі кездейсоқ таңдалады, ал осы элементке сәйкес келетін Aa_j деңгейлік жиынынан

x_i элементі де кездейсоқ таңдалады. Бұл жағдайда a_j таңдау механизмінің ықтималдығы, демек, Aa_j тең орнатылады

$$P(a_j) = P(Aa_j) = m(x_j) - m(x_{j-1}), \tag{7}$$

Яғни деңгейлер арасындағы интервалдың шамасымен анықталады. Aa_j – ден x_i элементін таңдау механизмінің тең деп қабылданады

$$P(x_i \mid Aa_j) = \begin{cases} 0, & x_i \in Aa_j \\ \dfrac{1}{n_j}, & x_i \notin Aa_j \end{cases} \tag{8}$$

Мұндағы n_j - Aa_j жиынының элементтерінің саны. $P(Aa_j)$, $P(x_i|Aa_j)$, j=1,2,…,n, ықтималдықтарын білу негізінде анықталған қаржылық көрсеткіштерге бақылау жүргізу және олардың күй класы туралы шешім қабылдау кезінде x_i элементін таңдаудың $P_i(x_i)$ ықтималдығы келесі формула бойынша есептелуі мүмкін:

$$P(x_i) = \sum_{j=1}^{n} P(x_i \mid Aa_j) P(a_j), \quad i = 1,2,...,n. \tag{9}$$

Жоғарыда (9) көрсетілген ережелерді ескере отырып, оған кіретін элементтерді табамыз:

$$\begin{aligned} &P(x_1) = \frac{1}{n} m(x_1); \quad P(x_2) = P(x_1) + \frac{1}{n-1}\left(m(x_2) - m(x_1)\right),...; \\ &P(x_{i+1}) = P(x_i) + \frac{1}{n-i}\left(m(x_{i+1}) - m(x_i)\right);... \\ &P(x_n) = P(x_{n-1}) + \left(m(x_n) - m(x_{n-1})\right). \end{aligned} \tag{10}$$

Бұл (10) теңдеулерді $m(x_i)$, i=1,2,…,n шамаларына қатысты шешуге болады:

$$\begin{cases} m(x_1) = nP(x_1); \quad m(x_2) = (n-1)P(x_2) + P(x_1);... \\ m(x_{i+1}) = (n-1)P(x_{i+1}) + \sum_{j=1}^{i} P(x_j); \quad m(x_n) = \sum_{j=1}^{n} P(x_j). \end{cases} \tag{11}$$

Ықтималдықтардың көрсетілген бағаларын алу және осылайша кәсіпорынның қаржылық жағдайын талдау кезінде тиесілілік дәрежесін анықтау үшін Р.Ягер процедурасының келесі модификациясын қолдану ұсынылады.

1. Әмбебап жиынның құрамына сәйкес $X=\{x_1, x_2,... ,x_n\}$ массивін қалыптастыру $\{(x_1,t_1), \{(x_2,t_2),\ldots \{(x_n,t_n)\}$ мұндағы t_i, $i=1,\ldots,n$ шамалары бастапқы нөлге тең деп есептеледі, қарастырылып отырған x_i пайда болу санына байланысты сарапшы экономисттен ақпарат жинауға арналған процедура. Қарастырылатын s деңгейлерінің санын таңдаңыз (әдетте $s>n$ таңдалады) және көптеген элементтерді орнатыңыз

$$\Omega=\left\{a_1=\frac{1}{s},\ a_2=\frac{2}{s},...,a_{s-1}=\frac{s-1}{s},\ a_s=1\right\}, \qquad (12)$$

Бір-бірінен тең деңгейлерді бекіту. Болашақта әр итерациялық циклде осы жиыннан бір элемент алынып тасталады.

2. $a_j \square \Omega$, элементін кездейсоқ таңдап, оны Ω жиыннан алып тастаңыз және сарапшы экономистке Aa_j деңгейлік жиынының құрамын анықтауды ұсыныңыз.

3. Әрбір $x_i \square Aa_j$ элемент үшін Aaj өрісі бұрынғы мәнге 1/nj мәнін қосу арқылы жаңа ti мәнін анықтаңыз, мұндағы n - Aa_j жиынының элементтерінің саны. Массивті Ω тексеріңіз; егер ол бос болмаса, 2-қадамға өтіңіз, ал егер ол бос болса, 4-қадамға өтіңіз. Для каждого элемента $x_i \square Aa_j$ определить новое значение t_i путем добавления к прежнему значению величины $1/n_j$, где n - число элементов множества. Проверить множество; если оно не пусто, перейти к шагу 2, а если оно пусто перейти- шагу 4.

4. Бағалауды анықтаңыз $P(x_i)=t_i\frac{1}{s},\ i=1,2,...,n$ және 5- қадамға өтіңіз.

5. Алынған ықтималдық бағаларын өсу бойынша реттеңіз (сәйкесінше нөмірлеуді өзгертіңіз) және, (11)-ге ауыстырып, қажетті тиістілік дәрежелерін анықтаңыз.

Жоғарыда келтірілген процедураның негізгі идеясы, сарапшы экономистке әр элемент үшін мүшелік дәрежесінің нақты мәндерін атаудың орнына, кездейсоқ таңдалған деңгейден жоғары орналасқан көптеген элементтердің құрамын көрсету ұсынылады.

Аталған алгоритмді іке асыру кезінде сарапшы экономист Aa_j жиындарын берудің дұрыстығын келесі шартының орындалуын тексеру арқылы бақылауға болады:

$$a_j < a_k \ \rightarrow\ Aa_j \subseteq Aa_k. \qquad (13)$$

Есептеу күрделілігі және масштабталуы

Әзірленген алгоритм екі сатылы процедура болып табылады:

1. р. Ягердің өзгертілген процедурасын қолдана отырып, тиістілік функцияларын қалыптастыру;

2. қаржылық индикаторлар кеңістігінде бұлыңғыр кластерлеу.

Ресми түрде, ұқсастық қатынасының өтпелі тұйықталуын және объектілердің толық жұптық сәйкестігін пайдаланған кезде, кластерлеудің есептеу күрделілігі N объектілерінің саны бойынша O(N2) ретінде өседі, бұл шекті жағдайда масштабтауды шектеуі мүмкін.

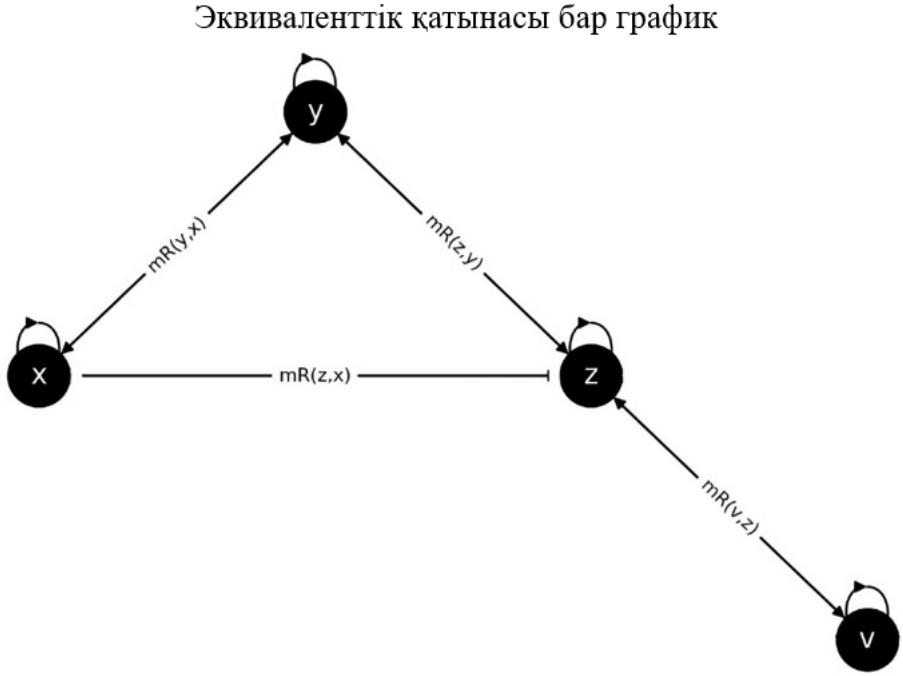
Практикалық іске асыруда ұқсастық матрицасын құру кезінде эвристикалық сүзу қолданылады: салыстыру қолтаңба кеңістігіндегі жақын көршілердің ішкі жиынымен шектеледі (k-NN-көршілестік), бұл кластерлеу сапасын айтарлықтай жоғалтпай жұптық операциялардың көлемін күрт азайтады. 1798 нысанның үлгісіндегі эксперименттік іске асыру (бір нысанға 35 белгі) стандартты жұмыс станциясында (Intel Core i7, 16 ГБ жедел жады; Python/NumPy-де іске асыру) 53 с толық екі сатылы процедураның орташа орындалу уақытын қамтамасыз етті. Жақын көршілерді іріктеу кезеңінің сызықтық масштабталуын ескере отырып, модель параметрлерін сақтай отырып, қолайлы жауап беру уақыты шеңберінде 10-20 мың объектіге дейінгі үлгілерді өңдеуге болады.

Осылайша, ұсынылған тәсіл толық емес деректерде талап етілетін интерпретация мен тұрақтылыққа ғана емес, сонымен қатар өнеркәсіптік сценарийлер үшін практикалық есептеу жарамдылығына ие (компания портфолиосын скрининг, банктік скоринг, мерзімді қайта есептеумен тәуекелдерді бақылау).

Кәсіпорындардың қаржылық жағдайын кластерлеу рәсімі

Бұлыңғыр математикадағы жіктеу есептерінің формальды орналасуы ұқсастық, ұқсастық (бұлыңғыр эквиваленттілік) қатынастарын және олардың арасындағы аралық позицияны алатын қатынастарды қолданумен байланысты (Mukhanov және т.б., 2023).

$r = <X, R>$ - соңғы жиынында берілген эквивалентіліктің (ұқсастықтың), бұлыңғыр қатынасы болсын. Берілген қатынастың графигінің кейбір, үш нүктеге салынған $x, y, z \in X$ ішкі графигін қарастырайық (сурет 1).



Сур. 1. Бұлыңғыр эквиваленттік қатынастың кіші графигі

Рефлексивтілік аксиомаларына байланысты шыңдар бірлікке тең дәрежесі бар ілмекке ие.

Симметрия аксиомасына сәйкес графиктің доғалары шарттарға бағынатын дәрежеге ие.

$$m_R(x,y)=m_R(y,x),\quad m_R(y,z)=m_R(z,y),\quad m_R(x,z)=m_R(z,x). \tag{14}$$

Өтпелі аксиоманың арқасында орын алады

$$\begin{aligned} m_R(x,y) &\geq \min\left(m_R(x,z),m_R(z,y)\right),\\ m_R(x,z) &\geq \min\left(m_R(x,y),m_R(y,z)\right),\\ m_R(y,z) &\geq \min\left(m_R(y,x),m_R(x,z)\right). \end{aligned} \tag{15}$$

Келтірілген аксиомалар негізінде ұқсастық қатынасының графы жұптарының тиістілік дәрежелері бағынатын маңызды қасиетті анықтауға болады. Болжайық , $m_R(x,y)\geq m_R(x,z)$. Онда транзитивтілік туралы соңғы қатынасқа сәйкес, мұнда деп есептейміз $m_R(y,x)=m_R(x,y)$, $m_R(y,z)\geq m_R(x,z)$, бар деп аламыз, ал екінші транзитивтілік қатынасы

бойынша - $m_R(x,z) \geq m_R(y,z)$. Сондықтан, қарастырылып отырған жағдайда $m_R(x,z) = m_R(y,z)$.

Егер біз бастапқы болжамды алға тартатын болсақ $m_R(x,y) \leq m_R(x,z)$, онда бізде бар соңғы тразитивті қатынасқа сәйкес $m_R(y,z) \geq m_R(x,y)$, онда бізде бар соңғы транзитивті қатынасқа сәйкес $m_R(z,y) = m_R(y,z)$, $m_R(x,y) \geq m_R(y,z)$, болады $m_R(x,y) = m_R(y,z)$.

Осылайша, кез келген *x, y, z* □ *X* нүктелерінің үштігі үшін ұқсастық қатынасының графы жұптарының тиістілік дәрежелері (ұқсастық графның доғалары) келесі заңға бағынатыны көрсетілді. Үш тиістілік дәрежелерінің кем дегенде екеуі бірдей, ал үшіншісі оларға тең немесе асып түседі. Осы заңнан келесі қорытындылар тікелей шығады. Егер аталған үштіктегі *x* нүктесі *y* нүктесімен байланысқан болса, онда *z* нүктесі не екі нүктемен байланысты, не біреуіне де байланысты емес. Сонымен қатар, егер *x*, *y*, *z* нүктелері байланысқан болса, $m_R(x,y) = m_R(y,x) > 0$, онда кез келген төртінші нүкте *v* (сурет.1) не осы нүктелердің әрқайсысымен байланысқан, не біреуімен де байланысқан емес. Бұл заңдылық *X* жиынына жаңа нүктелерді қосқан кезде де байқалады.

Барлық өзара байланысқан *X* жиынындағы нүктелер, мысалы, *x* □ *X* нүктесінің айналасында топтасатын нүктелер, ұқсастық қатынасының класының құрамын құрайды. *v* □ *X* нүктесін енгізе отырып, ол берілген класқа кірмеген жағдайда, жаңа ұқсастық қатынасы класын құруға болады және т.б. Нәтижесінде *X* жиыны өзара қиылыспайтын жиындар болып табылатын ұқсастық кластарға бөлінеді. Осылайша, алынған нәтиже кәдімгі эквиваленттік қатынасты енгізген кездегі нәтижеге ұқсас болып шығады. Мұндағы басты айырмашылық – әрбір кластағы элементтер арасындағы байланыстар жоғарыда шығарылған заңдылыққа бағытаны тиістілік дәрежелері арқылы бағаланады. Аталған тиістілік дәрежелерін пайдалану кәсіпорындардың экономикалық жағдайларының ұқсастық қатынасы ұғымына икемділік береді және оны қолдану үшін мүмкіндіктер жасайды (Mukhanov және т.б., 2023; Orazbayev және т.б., 2015).

[x_i] жиынын эквиваленттік класы ретінде сипаттау, кластың өкілі ретінде қай элемент x_i ді таңдағаныңызға қарамастан, бірдей болып қалады. Алайда, бұлдыр эквиваленттік (ұқсастық) қатынастар үшін кластың өкілі ретінде x_i элементін таңдау бұлдыр жиынды анықтайды. Бұл жиын *R* графында x_i мен жұп құрайтын элементтің жиынтығы ретінде қарастырылады және олардың осы жиынға тиістілік дәрежесі $m_R(x_i, x_j)$ ке тең. Мұндағы ерекшелік мынада: эквиваленттік қатынастың өкілдер класының әртүрлі элементтерін таңдау

элементтердің құрамы бірдей, бірақ жалпы жағдайда осы элементтердің тиістілік дәрежелері бойынша әртүрлі бұлдыр жиындарды тудырады.

Аталған бұлдыр жиындар эквиваленттік кластан айырмашылығы, ұқсастық прекластары деп аталады. Ұқсастық прекластарының саны $[x_i]_R$, $[x_j]_R$,... эквиваленттік класқа кіретін элементтер санына тең (кейбір ұқсастық прекластары бір-бірімен сәйкес келуі мүмкін). Кластерлік талдауда эквиваленттік қатынастар класы кластерлер ретінде қарастырылса, ал бұлдыр эквиваленттік қатынастардың прекластары (ұқсастық прекластары) – бұлдыр кластерлер болып табылады (Orazbayev және т.б., 2015; Pedrycz, 2015; Shipley және т.б., 2013). Эквиваленттік кластағы белгілі бір элементті xi және оған сәйкес бұлдыр кластерді [xi]R бөліп көрсету белгілі бір практикалық негіздемелерге байланысты болуы мүмкін. Мысалы, бұл элемент қаржылық құрылымда нақты мазмұнды интерпретацияға ие болуы мүмкін ().

Бұлдыр кластерлер бұлдыр қатынас матрицасы *r* -ді бергенде, оның жолдары немесе бағандары арқылы анықталады (әрбір жол немесе баған белгілі бір бұлдыр кластерге сәйкес келеді).

Берілген *P* ақырлы жиыны, ол *N* объектіден тұрады, $P=\{p^1, p^2,\ldots,p^N\}$, мұнда әрбір p^j объектісіне *m*-өлшемді кеңістіктегі ($z^j \in \mathrm{R}^m$) белгілі бір нақты нүкте z^j сәйкес келеді. Бұл кеңістік кәсіпорындардың қаржы ресурстарын қалыптастыру және пайдалану процесін бейнелейтін белгілер кеңістігі ретінде интерпретацияланады ().

P жиынын бұлдыр кластерлердің оңтайлы санына бұлдыр бөлуді табу қажет. Есепке нақтылық беру үшін, белгілер кеңістігінде *k* қосымша векторлық шамалар енгізіледі, $\vec{v}_i \in \mathrm{R}^m,\ i=1,2,...,k,$ олар кәсіпорынның қаржылық жағдайын сипаттайтын кластерлердің вектор-центроидтары ретінде интерпретацияланады. Нәтижесінде, m_{ij}, $\vec{v}_i$ үздіксіз айнымалыларымен сызықтық емес бағдарламалау есебі құрылады:

$$I=\sum_{i=1}^{k}\sum_{j=1}^{N} m_{ij}^2 \left\| \vec{z}^{\,j}-\vec{v}^{\,i} \right\|^2 \to \min_{\{m_{ij}\}\{v_i\}}. \tag{16}$$

$$\sum_{i=1}^{k} m_{ij}=1,\quad j=1,2,...,N,\quad \forall i,\forall j, m_{ij}\geq 0. \tag{17}$$

$$\vec{v}_i=\frac{1}{\sum_{j=1}^{N} m_{ij}^2}\sum_{j=1}^{N} m_{ij}^2 \vec{z}^{\,j},\quad i=1,2,...,k, \tag{18}$$

$\vec{v}_i$ - Ізделінетін *m*-өлшемді кеңістіктегі нүктелер, кластерді білдіретіндер;
m_{ij} - *j*-ші объектінің *i*-ші кластерге тиістілік дәрежесі;

$\left\| \vec{z}^{\,j} - \vec{v}^{\,i} \right\|$ - R^m кеңістігіндегі метрикті анықтайтын вектордың $\vec{z}^{\,j} - \vec{v}^{\,i}$ нормасы.

Осылайша, ұқсастық $r = <X, R>$ енгізу кластерлеу тапсырмасын екі алгоритмнің, яғни алдын ала және негізгі кластерлеудің кезектесіп қолданылуы арқылы тікелей шешуге мүмкіндік береді.

Алдын ала кластерлеу алгоритмі $P = \{p^1, p^2, \ldots, p^N\}$ объектілер жиынында ұқсастық қатынасын r енгізу және оның транзитивтілік тұйықталуын жүзеге асыру арқылы кластерлеуді жүргізуге бағытталған. Сонымен қатар, кәсіпорындардың қаржылық жағдайын сипаттайтын белгілер мәндері $Z = \{z^1, z^2, \ldots, z^N\}$, векторлары туралы деректер, тікелей $\bar{r}.$ -де көрсетілмесе де, экономист-эксперттердің ұқсастық туралы қорытындыларын қалыптастыру кезінде жанама түрде ескеріледі.

1. P объектілер жиынында ұқсастық қатынасын r құру:

2. Транзитивтілік тұйықталуын жүзеге асыру - ұқсастық қатынасын табу $\bar{r}.$

3. $\bar{r}$ кластерлеу негізінде P жиынын k эквиваленттілік кластерлеріне бөлу (k санын таңдау кәсіпорындардың қаржылық талдау тапсырмасын мазмұнды емес ресми түрде талдау негізінде жүзеге асырылады).

4. Әрбір k эквиваленттілік класы үшін r ұлдыр ұқсастық қатынасы негізінде сәйкес бұлдыр кластерлер жиыны анықталады, және соңынан осы жиыннан белгілі бір бұлдыр кластер таңдалады, мысалы, объектілердің тиістілік дәрежелерінің қосындысы максималды болатын кластер. Осыған байланысты, осы кластерге әрбір j - ші объектінің тиістілік дәрежелері m_{ij} бекітіледі.

5. Кәсіпорындардың қаржылық жағдайын сипаттайтын белгілер мәндерінің z^j векторлары және әрбір объектінің p, тиістілік дәрежелері m_{ij} негізінде (18) қатынасын пайдалана отырып, кластер $\vec{v}_i^{\,o}, \quad i = 1,2,...,k.$ орталықтарының нүктелері анықталады.

Келесі негізгі кластерлеу алгоритмі 3 қадамнан тұрады.

1. Белгілі m_{ij} мәндері бойынша (i=1,2,...,k; j=1,2,...,N) формуланы (18) -ді пайдалана отырып, орталықтарды анықтау қажет $\vec{v}_i, \quad i = 1,2,...,k.$. Итерациялық процесті бастаған кезде бұл қадам алдын ала кластерлеу алгоритмінің бесінші қадамымен (орталықтарды анықтаумен $\vec{v}_i^{\,o}, \quad i = 1,2,...,k$) сәйкес келеді.

2. Әрбір p^j объектісі үшін (j =1,2,...,N) формуланы пайдаланып анықтау қажет.

Нәтижелер және оларды талқылау

Тиістілік функцияларын алдын ала қалыптастырудың және

олардың негізінде кластерлеудің ұсынылған әдістемесі салалар бойынша кәсіпорындардың қаржылық жағдайын сипаттайтын (19) деректердің ашық үлгісі арқылы тексерілді (сурет 2, 3).

Эксперименттікзерттеулер35индикаторменсипатталғанпрецеденттердің бұлыңғыр кластерленуінен және кейіннен алынған кластерлердің құрылымын кәсіпорынның қалыпты/банкроттыққа дейінгі қаржылық жағдайының екі класын сипаттайтын нақты деректермен салыстырудан тұрды. Бұл тәсіл бұлыңғыр логика теориясына негізделген әдістеменің тиімділігін бағалауға және банкроттықтың алдын-ала болжамдарының дәлдігін тексеруге мүмкіндік берді.

$$m_{ij} = \frac{1}{\left(\mathbf{r}_{z_j} - \mathbf{r}_{v_i}\right)^2 \sum_{i=1}^{k} \frac{1}{\left\|\mathbf{r}_{z_j}, \mathbf{u}_{v'_i}\right\|^2}} = \frac{1}{\sum_{i=1}^{k} \left(\frac{\left\|\mathbf{r}_{z_j}, \mathbf{r}_{v_i}\right\|}{\left\|\mathbf{r}_{z_j}, \mathbf{u}_{v'_i}\right\|} \right)^2}, i = 1,...,k, j = 1,...,N \quad (9)$$

оның кластерлерге тиістілік дәрежесі m_{ij}, (i=1,2,...,k; j=1,2,...,N).

m_{ij} құндылықтарын есте сақтаңыз.

3. Егер барлық m_{ij} үшін шарт орындалса

$$\delta\left(m_{ij}, m'_{ij}\right) \le \varepsilon, \quad (20)$$

Онда процессі тоқтату қажет.

Керісінше жағдайда 1 қадамға оралу керек.

	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN
0	0.5	0.50	0.25	0.50	0.50	1.00	1.00	0.75	1.00	0.75	0.00	0.50
1	0.0	0.00	0.00	0.50	0.75	0.75	1.00	0.75	0.00	0.75	0.75	1.00
2	1.0	1.00	0.75	0.25	1.00	0.50	0.75	1.00	0.50	0.50	0.75	0.75
3	1.0	0.25	0.25	1.00	1.00	0.25	1.00	1.00	0.75	0.75	0.00	0.25
4	0.0	0.00	0.75	0.50	0.00	0.50	0.25	0.75	1.00	1.00	1.00	0.75

0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары...
1	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары...
2	банкрот
3	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары...
4	банкрот

Сур. 2. Қаржылық көрсеткіштер кестесінің ұйымдастырылуы

BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	B
0,5	0,5	0,25	0,5	0,5	1	1	0,75	1	0,75	0	0,5	0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0	0	0	0,5	0,75	0,75	1	0,75	0	0,75	0,75	1	0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
1	1	0,75	0,25	1	0,5	0,75	1	0,5	0,5	0,75	0,75	0	банкрот							
1	0,25	0,25	1	1	0,25	1	1	0,75	0,75	0	0,25	1	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0	0	0,75	0,5	0	0,5	0,25	0,75	1	1	1	0,75	1	банкрот							
0	1	1	0,5	1	1	0,25	0	0,75	0,25	0,75	0,5	1	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0,75	1	0,25	0,75	1	0,5	1	0,5	0,25	0,75	1	0	0	банкрот							
0,75	0,25	0,5	0,75	0,75	0,5	0,25	0,75	0,5	0,5	0	0,75	1	банкрот							
0,75	0,5	1	0,25	0,5	0,25	0,5	1	1	1	0,75	0,25	0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0,5	0,25	0,25	0,25	1	1	0,75	0,5	0	0,5	1	0,5	0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы төмен							
0,5	0,25	1	0,25	0,25	1	0,25	0,75	0,5	0,25	0,75	0,75	0	банкрот							
1	0,75	1	0,5	0,25	0,25	0,5	0,25	0,5	0,5	0,25	0,5	0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0,5	0,5	1	0,75	1	0,5	0,75	0,25	1	0,75	0	0,5	1	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0,75	0,75	0,5	0,75	1	0,25	0,25	0,25	1	1	1	0	0	банкрот							
0	0,5	0,5	1	1	0	0,75	0,75	1	0,75	0,25	1	1	банкрот							
0	1	0,25	0,5	0,75	2	1	0,25	0,5	0,75	0	1	0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0,5	0,5	0	0,5	0,25	0,25	0	0,25	0	0,5	1	0,75	1	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0	0,5	1	0,5	0,75	1	1	0	0,25	0	0,75	0	1	банкрот							
0,5	0,25	0,5	0,25	1	0	0,5	1	1	0,75	1	0,75	0	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы жоғары							
0,5	1	0,75	0,25	0,75	0,25	1	0,25	0,5	1	0,75	0	1	бұл кәсіпорынның банкроттыққа ұшырау ықтималдығы төмн							

Сур. 3. Қаржылық көрсеткіштер кестесінің құрылымы

Мұғалімнің нақты қалыпты немесе банкроттыққа дейінгі қаржылық жағдайларға қатысты нұсқаулары 11 негізгі параметр түрінде ұсынылды (сурет 4). Бұл параметрлер қаржылық тұрақтылық, өтімділік, активтерді басқару тиімділігі, рентабельдік коэффициенттері, қарыз жүктемесі және басқа да көрсеткіштерді қамтыды. Осы индикаторлар банкроттықтың ықтималдығын бағалауда шешуші рөл атқарды және кластерлеудің тиімділігін жақсарту үшін қолданылды.

Сурет 4 бұлыңғыр логикаға негізделген әдістеменің банкроттықтың шынайы деректерімен сәйкес келу дәлдігіне қалай әсер ететінін көрсетеді. Бұл зерттеу барысында индикаторлар санының өзгеруі әдістеменің сенімділігіне қалай ықпал ететіні бағаланды. Нәтижелер көрсеткендей, банкроттықты болжау дәлдігі қолданылатын индикаторлар санының өсуімен артады, алайда белгілі бір шекке жеткенде артық параметрлерді қосу нәтиженің айтарлықтай жақсаруына әкелмейді.

Танудың сенімділігін анықтау әдісі дұрыс болжамдардың санын, яғни банкроттықтың нақты бақыланатын кластарымен сәйкес келетін жауаптар санын есептеуге және жиілік көрсеткішін анықтауға негізделген. Бұл тәсіл кластерлеу нәтижелерінің объективтілігін бағалауға, сондай-ақ әдістеменің дәстүрлі статистикалық әдістермен салыстырғанда қаншалықты тиімді екенін көрсетуге мүмкіндік берді.

Эксперименттік зерттеулер барысында нейрондық желілер және басқа да машиналық оқыту әдістерімен салыстыру жүргізілді. Бұлыңғыр логикаға негізделген әдістеменің банкроттықты болжаудағы дәлдігі машиналық оқыту әдістерімен салыстырғанда айтарлықтай жоғары болған жоқ, бірақ оның интерпретациялауға ыңғайлы болуы және сараптамалық бағалауларды қоса алу мүмкіндігі оны қаржылық талдау үшін тиімді құрал етеді.

Осылайша, эксперименттік зерттеулер бұлыңғыр кластерлеу негізіндегі әдістеменің банкроттықтың ықтималдығын бағалаудағы тиімділігін дәлелдеді, сондай-ақ қаржылық көрсеткіштерді талдауда сараптамалық тәсілді қолданудың артықшылықтарын көрсетті. Бұл әдіс қаржылық болжау жүйелерін жетілдірудің перспективалы бағыты ретінде қарастырылуы мүмкін:

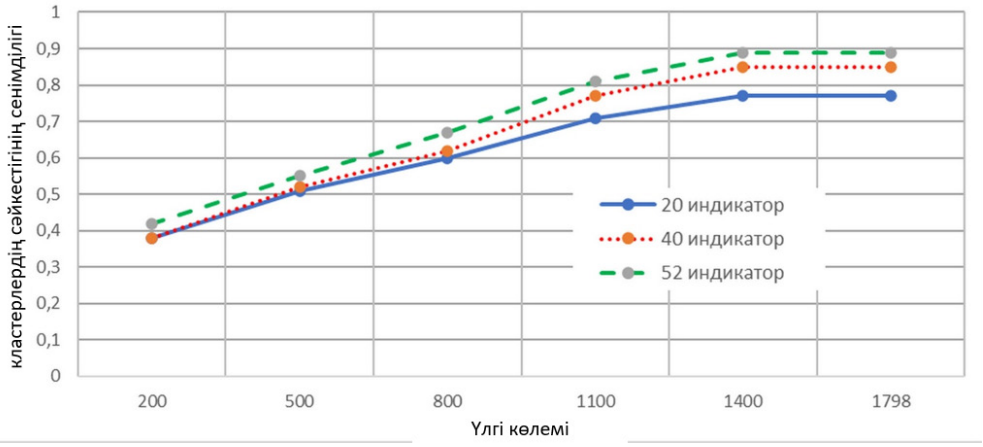
$$D = \frac{N_1}{N}, \tag{21}$$

N_1 – кластерге дұрыс жатқызылғандардың саны, N – екі кластердің біреуіне жатқызылғандардың жалпы саны.

x1	Аймақ саласындағы бәсекелестік деңгейі
x2	Реглондағы салалық тәуекелдердің шоғырлануы
x3	ЖӨӨ-де кәсіпорын жұмыс істейтін саланың үлесі.
x4	Аймақтағы экономикалық жағдай
x5	Аймақтық және макроэкономикалық тәуекелдердің әсері.
x6	Несие тарихының сапасы (болуы/болмауы)
x7	Кәсіпорынған алынған кредиттер (қарыздар) саны
x8	Кепіл.
x9	Аймақтағы кәсіпорынның қызмет ету мерзімі
x10	Қолма-қол жабдықтар паркінің коэффициенті
x11	Орнатылған жабдықтар паркінің коэффициенті
x12	Жабдықты жүктеу қаркындылығы коэффициенті
x13	Жоспардың орындалу коэффициенті (жалпы өнім бойынша)
x14	Сала бойынша кәсіпорынның орташа сатылатын көбіктерін бағалау.
x15	Жалпы шығарылымдағы жаңа өнімнің (қызметтің) үлес салмағы
x16	Өнімнің(қызметтің)өзіндік құнындағы ірі жеткізушінің үлесі
x17	Өнімді сатудан түскен Түсімдегі ірі сатып алушының үлесі
x18	Ең еңбекке қабілетті және білікті персоналдың үлес салмағы
x19	Кадрлармен қамтамасыз етудің жалпы коэффициенті
x20	Кәсіпорын қызметкерлерінің біліктілік деңгейі
x21	Біліктіліктің есу индексі
x22	Қаржы менеджменті саласындағы кадр құрамының біліктілігі
x23	Айналым коэффициенті.
x24	Кейтцтің кері коэффициенті
x25	Ағымдағы өтімділік коэффициенті
x26	Меншікті айналым қаражатымен қамтамасыз ету коэффициенті
x27	Дебиторлық берешек айналымы кезеңі
x28	Кредиторлық берешек айналымы кезеңі
x29	Берешек айналымы кезеңдерінің қатынасы.
x30	Дайын өнім қалдықтарының айналым кезеңі!
x31	Қаржылық белсенділік коэффициенті
x32	Қаржылық тәуелсіздік коэффициенті.
x33	Материалдық қорларды өз қаражатымен қамтамасыз ету коэффициенті:
x34	Жалпы рентабельділік
x35	Сатудың кірістілік коэффициенті.

REG	Аймақтық және салалық ерекшеліктерді бағалау.
KREDIT	Несиелік бағалау
THE	Техникалық жарақтандыруды бағалау
MARKET	Нарықтық әлеуетті бағалау
STAFF	Кадрлық қамтамасыз ету сапасы
PSICH	Кәсіпорындағы моральдық-психологиялық ахуалды бағалау
ABILITY	Өтімділікті бағалау
TURN	Айналым көрсеткіштерін бағалау
FINN	Қаржылық тұрақтылықты бағалау
Z25	25 қаржылық тұрақтылықтың интегралдық көрсеткіші
Z35	айнымалылар
	35 қаржылық тұрақтылықтың интегралдық көрсеткіші
	айнымалылар

Сур. 4 . Қаржы жағдайларының 1798 мысалынан тұратын оқу жиынын 6 сала бойынша бөлудің схемасы. Бұл схема көрсеткіштер кеңістігінде кластерлеу мен 2 жағдайды сипаттайтын танылатын параметрлерді жүзеге асыруды қамтиды.



Сур. 5. Даму әдістемесі бойынша құрылған кластерлердің және банкроттықтың нақты деректерінің сәйкестік дәлдігіне тәуелділігі

Дамыған әдістеме бойынша кластерлерге жатқызу сапасы оқыту жиынының көлеміне айтарлықтай тәуелді екенін көрсетеді және дәлдік деңгейлері бойынша кәсіпорындардың қаржылық жағдайын болжауда нейрондық желі әдістері мен алгоритмдік дихотомиялармен салыстыруға болады.

Нейрондық желілік тәсілдермен салыстырғанда күйді тану тиімділігін тексеру үшін классикалық көп қабатты перцептрон архитектурасын (Multilayer perceptron, MLP) қолдану арқылы эксперимент жүргізілді. Модельде сәйкесінше 64 және 32 нейроннан тұратын екі жасырын қабат қолданылды, белсендіру функциясы-ReLU, Шығыс қабаты - сигмоидты. Оқыту binary cross-entropy жоғалту функциясымен 100 ғасыр бойы Adam алгоритмін қолдана отырып жүргізілді. Оқыту және тестілеу деректері анық емес әдістемедегідей болды: экономиканың 6 саласын қамтитын кәсіпорындардың қаржылық жағдайының 1798 прецедентінің үлгісі, әрбір жазбада 35 қалыпқа келтірілген қаржылық белгілер болды. Оқыту және тестілеу үлгілеріне бөлу 80/20 пропорциясында жүзеге асырылды.

Нәтижелер MLP «норма» / «банкроттық» күйлерін жіктеу кезінде 91,2 % дәлдікке жеткенін көрсетті. Тиістілік функцияларын қалыптастыруға және кейіннен бұлыңғыр кластерлеуге негізделген ұсынылған әдіс тиімділіктің салыстырмалы деңгейін көрсететін 89,4 % дәлдікке жетті. Сонымен қатар, ұсынылған әдістеме маңызды артықшылықтарға ие: деректердің толық устойчивстігіне төзімділік, тиістіліктің семантикалық функцияларын қалыптастыруға сарапшылардың қатысуы есебінен шешімдердің түсіндірілуі, сондай-ақәртүрлісараптамалықсценарийлергебейімделумүмкіндігі.Осылайша, MLP типтік архитектурасымен салыстыру бұлыңғыр логикаға негізделген әдіс кәсіпорындардың қаржылық тұрақтылығын болжау міндеттерінде нейрондық желілік тәсілдермен бәсекеге қабілетті деген қорытындыға келеді.

Қорытынды

Әрбір әлемдік қаржы дағдарысы көптеген елдерде банкроттық санының артуына әкеледі, бұл осы класстағы оқиғаларды болжау құралдарының сенімділігін арттыру мен әртараптандыруға деген қызығушылықты күшейтеді. Кәсіпорындардың қаржылық жағдайының белгісіздігі мәселесін терең зерттеу, бұл құбылыстың көпқырлылығын көрсетеді және негізгі экономикалық санаттар мен қазіргі заманғы математикалық үлгілерге негізделген әдістерді қолдану қажеттілігін айқындайды.

Қаржылық жағдайды талдау мен болжауда ықтималдық математиканың, бұлыңғыр логика мен статистикалық әдістердің кешенді түрде қолданылуы жоғары дәлдікті нәтижелерге қол жеткізуге мүмкіндік береді. Бұл тәсіл экономикалық белгісіздіктерді, ақпараттың толық еместігін және қаржылық көрсеткіштердің өзгермелілігін ескере отырып, дәлірек талдау жасауға мүмкіндік береді.

Зерттеу барысында қаржылық индикаторлар кеңістігінде тиесілілік

функцияларын синтездеу және кластерлеу процедурасын қамтитын жаңа әдістеме әзірленді. Бұл әдістеме бұлыңғыр логиканың классикалық артықшылықтарын іске асыруға мүмкіндік береді және оны банкроттықты болжау мен қаржылық жағдайды бағалау құралдарының бірі ретінде қарастыруға болады.

Ұсынылған тәсіл қарама-қайшы, толық емес деректер мен белгісіздік факторлары жағдайында қанағаттанарлық нәтижелер алуға мүмкіндік береді. Сонымен қатар, сарапшылардың тәжірибесін есепке ала отырып, бұл әдіс интуитивті түрде түсінікті және қолдануға ыңғайлы құралға айналады.

Зерттеу нәтижелері көрсеткендей, бұлыңғыр логикаға негізделген талдау әдістемесі дәстүрлі статистикалық тәсілдермен салыстырғанда жоғары дәлдікке ие және шешім қабылдауда икемділікті арттыруға мүмкіндік береді. Бұл әдістеме қаржылық диагностика, бизнес-аналитика және стратегиялық жоспарлау салаларында қолдануға жарамды, сондай-ақ кәсіпорындардың банкроттық қаупін бағалау үшін тиімді құрал бола алады.

Алдағы зерттеулерде әдістеменің басқа математикалық модельдермен үйлесімділігін арттыру, сондай-ақ динамикалық нарықтық жағдайларға бейімдеу мәселелерін қарастыру перспективалы бағыттардың бірі болып табылады.

REFERENCES

Абдукаримов И.Т., Беспалов М.В. (2023). Анализ финансового состояния и финансовых результатов предпринимательских структур. — М.: ИНФРА-М. — 214 с.

Анализ финансового состояния предприятия. Оценка вероятности банкротства: Модель Альтмана (Z-модель) (Электронный ресурс). — URL: http://www.afdanalyse.ru/publ/bankrostvo/bankrot_1/13-1-0-10 (дата обращения: 16.06.2024).

Анализ финансового состояния предприятия. Оценка вероятности банкротства: Модель Таффлера и Тишоу (Электронный ресурс). — URL: http://afdanalyse.ru/publ/finansovyj_analiz/1/bankrot_tafler/13-1-0-37 (дата обращения: 16.06.2024).

Анализ финансовой устойчивости: сайт Audit-it.ru (финансовый анализ по данным отчетности) (Электронный ресурс). — URL: https://www.audit-it.ru/finanaliz/terms/solvency/ (дата обращения: 15.06.2024).

Altman E., Hotchkiss E. (2006). Corporate Financial Distress and Bankruptcy: Predict and Avoid Bankruptcy, Analyze and Invest in Distressed Debt. — John Wiley and Sons, Ltd. — 368 p.

Bernstein L.A. (Бернстайн Л.А.) (2015). Анализ финансовой отчетности: теория, практика и интерпретация. — М.: Финансы и статистика. — 624 с.

Carvalho J.P., Tome J.A.B. (2009). Rule based fuzzy cognitive maps in socio-economic systems // IFSA–EUSFLAT 2009 Proceedings. — Lisbon. — Pp. 1821–1826.

Ефанова Н.В., Ващенко В.Р. (2019). Использование методов нечеткой логики в оценке финансовой устойчивости предприятия // Вестник Воронежского государственного аграрного университета. — 2019. — № 3 (62). — С. 206–212.

Helfert E. (Хелферт Э.) (2015). Техника финансового анализа. — М.: Аудит, Юнити. — 663 с.

Muhd Khairulzaman Abdul Kadir, Hines E.L., Qaddoum K., Collier R., Dowler E., Grant W., Leeson M., Iliescu D., Subramanian A., Richards K., Merali Y., Napier R. (2013). Food security risk level assessment: a fuzzy logic-based approach // Applied Artificial Intelligence. — 2013. — Vol. 27. — No. 1. — Pp. 50–61. DOI: 10.1080/08839514.2013.747372.

Mukhanov S.B., Alimbekov A.Ye., Marat G.S., Uatbayeva A.M., Aldanazar A.A. (2021). Automation of staff recruitment and assessment // Международный журнал информационных и коммуникационных технологий. — 2021. — 1(3). DOI: 10.54309/IJICT.2020.3.3.009.

Mukhanov S., Uskenbayeva R., Young I.C., Yemberdiyeva A., Bekaulova Z. (2024). Deep and machine learning models for recognizing static and dynamic gestures of the Kazakh alphabet // Scientific Journal of Astana IT

University. — 2024. — No. 18. — Pp. 75–95. DOI: 10.37943/18JYLU4904.

Mukhanov S., Uskenbayeva R., Young Im Cho, Kabyl D., Les N., Amangeldi M. (2023). Gesture recognition of machine learning and convolutional neural network methods for Kazakh sign language // Scientific Journal of Astana IT University. — 2023. — 15(15). — Pp. 85–100. DOI: 10.37943/15LPCU4095.

Mukhanov S., Uskenbayeva R., Rakhim A., Akim A., Mamanova S. (2024). Gesture recognition of the Kazakh alphabet based on machine and deep learning models // Procedia Computer Science. — 2024. — Vol. 241. — Pp. 458–463. DOI: 10.1016/j.procs.2024.08.064.

Mukhanov S.B., Lee A.S., Zheksenov D.B., Yevdokimov D.D., Amirgaliev E.N., Kalzhigitov N.K. (2023). Comparative analysis of neural network models for gesture recognition methods of hands // Bulletin of NIA RK. Information and communication technologies. — 2023. — No. 2. — P. 88.

Недосекин А.О., Овсянко А.Н. (б.г.). Нечетко-множественный подход в маркетинговых исследованиях (Электронный ресурс). — URL: http://www.aup.ru/articles/marketing/15.htm; http://www.vmgroup.sp.ru (дата обращения: 16.06.2024).

Orazbayev B.B., Orazbayeva K.N., Kurmangaziyeva L.T., Makhatova V.E. (2015). Multi-criteria optimisation problems for chemical engineering systems and algorithms for their solution based on fuzzy mathematical methods // EXCLI Journal. — 2015. — Vol. 14. — Pp. 984–998.

Pedrycz W. (2015). From fuzzy data analysis and fuzzy regression to granular fuzzy data analysis // Fuzzy Sets and Systems. — 2015. — Vol. 274. — Pp. 12–17.

Shipley M.F., Johnson M., Pointer L., Yankov N. (2013). A fuzzy attractiveness of market entry (FAME) model for market selection decisions // The Journal of the Operational Research Society. — 2013. — Vol. 64. — No. 4. — Pp. 597–610.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 136–152
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.008

ӘОЖ 629.78: 004.27
UDC 629.78: 004.27

DEVELOPMENT AND BENCH TESTING OF AN ONBOARD CONTROL COMPLEX FOR CUBESAT-CLASS NANOSATELLITES

A. Boranbayeva [1,2,3*], G. Serghazin[4], Y. Nurgizat[5], K. Bagitova[6], T. Iliev [7]
[1]Al-Farabi Kazakh National University, Kazakhstan, Almaty;
[2]Academician U. Joldasbekov Institute of Mechanics and Machine Science, Kazakhstan, Almaty;
[3]Satbayev Kazakh National Technical Research University, Almaty, Kazakhstan;
[4]M. Tynyshbaev ALT University, Almaty, Kazakhstan;
[5]G. Daukeyev Almaty University of Power Engineering and Telecommunications, Almaty, Kazakhstan;
[6]Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan;
[7]Angel Kanchev University of Ruse, Ruse, Bulgaria.
E-mail: boranbayeva.anargul.talgatkyzy@gmail.com

Anargul Boranbayeva — 3rd-year PhD student, Al-Farabi Kazakh National University, researcher, the Institute of Mechanics and Machine Science named after Academician O.A. Zholdasbekov, and senior lecturer, Satbayev Kazakh National Technical Research University, Almaty, Kazakhstan
E-mail:boranbayeva.anargul.talgatkyzy@gmail.com,
https://orcid.org/0000-0003-2221-9604;
Gani Serghazin — PhD, associate professor, ALT University, Almaty, Kazakhstan
https://orcid.org/0000-0003-2762-473;
Yerkebulan Nurgizat — PhD, G. Daukeyev Almaty University of Power Engineering and Telecommunications, Almaty, Kazakhstan
https://orcid.org/0000-0002-9712-5592;
Kalamkas Bagitova — PhD, associate professor, Head of the Computer Science Department, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan
https://orcid.org/0000-0003-1587-1995;
Teodor Iliev — PhD, professor, University of Ruse, Ruse, Bulgaria.
https://orcid.org/0000-0003-2214-8092.

Abstract. This work presents the development of an onboard control complex

(OBC) for CubeSat-class nanosatellites with emphasis on energy safety, verifiability, and fault tolerance. The architecture follows a "space–ground" functional split: on board, deterministic protections (threshold/hysteresis), state-of-charge (SoC) estimation via an Extended Kalman Filter, and a schedule executor run on an STM32H743 microcontroller supported by an external watchdog timer; on the ground, energy-balance forecasting and schedule optimization account for illumination and communication windows. The sensor suite leverages accessible COTS components (IMU, magnetometer, barometer), while attitude estimation combines point solutions (TRIAD/QUEST) with recursive refinement (MEKF/EKF). A two-tier storage subsystem-QSPI-NAND for critical logs and SD via SDIO for bulk telemetry-prioritizes writes and remains robust under power shortfalls. Bench tests confirmed the correctness of the start-up state machine, stable sensor I/O, and preservation of the "black box" under energy deficit. Planned next steps include finalizing the firmware, introducing a lightweight AI layer (primarily groundside), extended climatic/vibrodynamic and reliability testing, and cybersecurity hardening. The resulting platform provides a basis for practical deployment in 1U–3U missions with stringent mass and power budgets.

Keywords: nanosatellite, onboard control complex, attitude and navigation system, inertial-magnetic navigation, attitude estimation, microcontrollers, energy efficiency and power management

For citation: A. Boranbayeva, G. Serghazin, Y. Nurgizat, K. Bagitova, T. Iliev. Development and Bench Testing of an Onboard Control Complex for Cubesat-Class Nanosatellites// International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 136–152. (In Russ.). https://doi.org/10.54309/IJICT.2025.24.4.008.

Conflict of interest: The authors declare that there is no conflict of interest.

CUBESAT ПІШІНДЕГІ НАНОСПУТНИКТЕР ҮШІН БОРТТЫҚ БАСҚАРУ КЕШЕНІН ДАЙЫНДАУ ЖӘНЕ СТЕНДТІК ТЕКСЕРУ

А.Т. Боранбаева [1,2,3]*, Ғ.Қ. Серғазин[4] , Е.С. Нұргизат[5], Қ.Б. Багитова[6], Т. Илиев[7]

[1]Әл Фараби атындағы Қазақ Ұлттық университеті, Алматы, Қазақстан;
[2]Академик Ө.А. Жолдасбеков атындағы механика және машинатану институты, Алматы, Қазақстан;
[3]Қ.И. Сәтпаев атындағы Қазақ Ұлттық техникалық зерттеу университеті, Алматы, Қазақстан;
[4]М. Тынышбаев атындағы ALT университеті, Алматы, Қазақстан;
[5]Ғ. Дәукеев атындағы Алматы энергетика және байланыс университеті, Алматы, Қазақстан;
[6]Х.Досмұхамедов атындағы Атырау университеті, Алматы, Қазақстан;
[7]Ангел Кынчев атындағы Русе университеті, Русе, Болгария.
E-mail: boranbayeva.anargul.talgatkyzy@gmail.com

Боранбаева Анаргүл — Әл Фараби атындағы Қазақ Ұлттық Университетінің 3-ші курс докторанты, Академик Ө.А. Жолдасбеков атындағы механика және машинатану институтының ғылыми қызметкері, Қ.И. Сәтпаев атындағы Қазақ Ұлттық техникалық зерттеу университетінің аға оқытушысы, Алматы, Қазақстан
E-mail: boranbayeva.anargul.talgatkyzy@gmail.com
https://orcid.org/0000-0003-2221-9604;
Серғазин Ғани — PhD, қауымдастырылған профессор, ALT университеті, Алматы, Қазақстан
https://orcid.org/0000-0003-2762-473;
Нұрғизат Еркебұлан — PhD, Ғ. Дәукеев атындағы Алматы энергетика және байланыс университеті, Алматы, Қазақстан
https://orcid.org/0000-0002-9712-5592;
Багитова Қаламқас — PhD, қауымдастырылған профессор, Х. Досмұхамедов атындағы Атырау университетінің «Информатика» кафедрасының меңгерушісі, Атырау, Қазақстан
https://orcid.org/0000-0003-1587-1995;
Илиев Теодор — PhD, профессор, Русе университеті, Болгария
https://orcid.org/0000-0003-2214-8092.

Аннотация. Жұмыс CubeSat форматына арналған борттық басқару кешенін (ББК) әзірлеуге арналған және энергетикалық қауіпсіздік, верификациялануғыштық пен ақауға төзімділікке басымдық береді. Архитектура «борт–жер» функцияларын бөлу қағидатына негізделген: бортта шекті/гистерезистік қорғаныстар, кеңейтілген Калман сүзгісі негізіндегі аккумулятордың зарядталу дәрежесін (SoC) бағалау және STM32H743 микроконтроллерінде сыртқы «watchdog»-таймермен орындалатын жоспарлар іске асырылған; жердегі контурда жарықтану мен байланыс терезелерін ескере отырып, энергия балансын болжау және жоспарларды оңтайландыру орындалады. Сенсорлық контур қолжетімді COTS-компоненттеріне (IMU, магнитометр, барометр) негізделген, ал бағдарды бағалау TRIAD/QUEST сияқты нүктелік алгоритмдерді MEKF/EKF рекурсивті әдістерімен ұштастырады. Екі деңгейлі деректер сақтау қосалқы жүйесі (сыни журналдар үшін QSPI-NAND және жаппай телеметрия үшін SDIO арқылы SD) жазуларды басымдықпен жүргізуді және энергия тапшылығы кезінде орнықтылықты қамтамасыз етеді. Стендтік сынақтар іске қосу автоматының дұрыстығын, датчиктермен тұрақты алмасуды және энергия тапшылығы жағдайында «қара жәшіктің» сақталуын растады. Сонымен қатар, микробағдарламаны аяқтау, жеңіл жасанды интеллект қабатын (негізінен жердегі жағында) енгізу, кеңейтілген климаттық-вибродинамикалық және сенімділік сынақтары, сондай-ақ киберқауіпсіздік шаралары жоспарланып

отыр. Ұсынылған платформа масса және энергия бюджеті қатаң шектелген 1U-3U миссияларында практикалық қолдануға негіз қалайды.

Түйін сөздер: наноспутник, борттық басқару кешені, бағдарлау және навигация жүйесі, инерциалды-магниттік навигация, бағдарды бағалау, микроконтроллерлер, энергия тиімділігі және энергияны басқару

Дәйексөздер үшін: Т. Боранбаева, Ғ.Қ. Серғазин, Е.С. Нұрғизат, Қ.Б. Багитова, Т. Илиев. Cubesat пішіндегі наноспутниктер үшін борттық басқару кешенін дайындау және стендтік тексеру// Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 136–152 бет. (Орыс тілі). https://doi.org/10.54309/IJICT.2025.24.4.008.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА И СТЕНДОВАЯ ВЕРИФИКАЦИЯ БОРТОВОГО КОМПЛЕКСА УПРАВЛЕНИЯ ДЛЯ НАНОСПУТНИКОВ ФОРМАТА CUBESAT

А.Т. Боранбаева [1,2,3*], Г.К. Сергазин[4] , Е.С. Нургизат[5], К.Б. Багитова[6], Т. Илиев[7]

[1]Казахский национальный университет им. аль-Фараби, Алматы, Казахстан;
[2]Институт механики и машиноведения им. академика У.А. Джолдасбекова, Алматы, Казахстан;

[3]Казахский национальный технический исследовательский университет им. К.И. Сатпаева, Алматы, Казахстан;

[4]Университет ALT имени М. Тынышбаева, Алматы, Казахстан;
[5]Алматинский университет энергетики и связи им. Г. Даукеева, Алматы, Казахстан;

[6]Атырауский университет им. Х. Досмухамедова, Атырау, Казахстан;
[7]Русенский университет им. Ангела Кынчева, Русе, Болгария.
E-mail: boranbayeva.anargul.talgatkyzy@gmail.com

Боранбаева Анаргуль — докторант 3-го курса, Казахский национальный университет имени аль-Фараби, научный сотрудник, Институт механики и машиноведения имени академика О.А. Джолдасбекова, старший преподаватель, Казахский национальный технический исследовательский университет имени К.И. Сатпаева, Алматы, Казахстан
E-mail:boranbayeva.anargul.talgatkyzy@gmail.com,
https://orcid.org/0000-0003-2221-9604;
Сергазин Гани — PhD, ассоциированный профессор, университет ALT, Алматы, Казахстан
https://orcid.org/0000-0003-2762-473;
Нургизат Еркебулан — PhD, Алматинский университет энергетики и связи

имени Гумарбека Даукеева, Алматы, Казахстан
https://orcid.org/0000-0002-9712-5592;
Багитова Каламкас — PhD, ассоциированный профессор, заведующая кафедрой информатики, Атырауский университет имени Халела Досмухамедова, Атырау, Казахстан
https://orcid.org/0000-0003-1587-1995;
Илиев Теодор — PhD, профессор, Русенский университет, Русе, Болгария
https://orcid.org/0000-0003-2214-8092.

Аннотация. Работа посвящена разработке бортового комплекса управления (БКУ) для наноспутников формата CubeSat с акцентом на энергетическую безопасность, верифицируемость и отказоустойчивость. Предложена архитектура с разделением функций между бортом и наземным контуром: на борту реализованы детерминированные механизмы защиты (порог/гистерезис), оценка степени заряженности аккумулятора на основе расширенного фильтра Калмана и исполнитель расписаний на микроконтроллере STM32H743 с внешним сторожевым таймером; на земле - прогноз энергетического баланса и оптимизация расписаний с учётом освещённости и окон связи. Сенсорный контур построен на доступных COTS-компонентах (IMU, магнитометр, барометр), а оценка ориентации комбинирует точечные алгоритмы (TRIAD/QUEST) с рекурсивным уточнением (MEKF/EKF). Двухуровневая подсистема хранения (QSPI-NAND для критических журналов и SD по SDIO для массивной телеметрии) поддерживает приоритизацию записей и устойчивость к энергетическим просадкам. Стендовые испытания подтвердили корректность конечного автомата запуска, стабильный обмен с датчиками и сохранность «чёрного ящика» при дефиците энергии. Обсуждаются планируемые шаги по завершению прошивки, внедрению лёгкой ИИ-прослойки (преимущественно на наземной стороне), расширенным климато-вибродинамическим и надёжностным тестам, а также кибербезопасности. Представленная платформа формирует основу для практического применения в миссиях 1U-3U с жёсткими ограничениями по массе и энергобалансу.

Ключевые слова: наноспутник, бортовой комплекс управления, система ориентации и навигации, инерциально-магнитная навигация, оценка ориентации, микроконтроллеры, энергоэффективность и управление энергией

Для цитирования: Т. Боранбаева, Ғ.Қ. Серғазин, Е.С. Нұрғизат, Қ.Б. Багитова, Т. Илиев. Разработка и стендовая верификация бортового комплекса управления для наноспутников формата cubesat//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 23. Стр. 136–152. (На русс.). https://doi.org/10.54309/IJICT.2025.24.4.008.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

За последние два десятилетия формат CubeSat стал одним из ключевых драйверов «демократизации» космической деятельности: стандартизация механики и интерфейсов, удешевление микроэлектроники и расширение доступа к пускам позволили университетам, НИИ и технологическим компаниям реализовывать научные и прикладные миссии в сжатые сроки и с умеренными бюджетами (Poghosyan et al., 2017; Liddle et al., 2020; Nieto-Peroy et al., 2019). При этом специфические факторы низких околоземных орбит (LEO) - неоднородности и вариации плотности термосферы, аэродинамические возмущения, гравитационные и магнитные поля - формируют повышенные угловые ускорения и, как следствие, сокращённый срок активного существования малых космических аппаратов (КА) по сравнению с аппаратами большого класса (Belokonov et al., 2018). Масштабирование числа запусков и переход к созвездиям усиливают требования к предсказуемости энергетики, устойчивости к отказам и проверяемости критических функций бортовой аппаратуры уже на этапах проектирования и стендовой отработки (Rigo et al., 2022).

Для аппаратов формата 1U-3U бортовой комплекс управления (БКУ) должен, как минимум, обеспечивать: (i) надёжный старт/перезапуск и базовую отказоустойчивость, (ii) устойчивое определение и управление ориентацией на базе бюджетных датчиков, (iii) энерго-безопасное планирование включений подсистем при ограниченном энергобалансе, (iv) детерминированность и верифицируемость алгоритмов с учётом ограниченной радиационной стойкости COTS-элементной базы. Классический «онборд-центристский» подход, когда все интеллектуальные функции переносятся на борт, повышает вычислительные и энергетические риски и усложняет верификацию. В качестве альтернативы требуется архитектура с жёстко детерминированной, проверяемой бортовой логикой и вынесением ресурсоёмких задач в наземный контур (Finance et al., 2021).

В данной работе предложен БКУ для CubeSat, реализующий указанную парадигму разделения: на борту - порогово-гистерезисные защиты, оценка степени заряженности аккумулятора (SoC) на базе расширенного фильтра Калмана и исполнитель расписаний на микроконтроллере STM32H743 в сочетании с внешним сторожевым таймером; на земле — прогноз энергетического баланса и оптимизация расписаний с учётом освещённости и окон связи. Навигация строится на совмещении бюджетных сенсоров (IMU LSM6DSV32, магнитометр/акселерометр LSM303AGR, резервный барометр BMP585 для стендовой калибровки и гермоконтроля), а оценка ориентации - на детерминированных алгоритмах (TRIAD/QUEST) с рекурсивным уточнением (MEKF/EKF), что соответствует современной практике малых КА (Finance et al., 2021; Dos Santos et al., 2021) Двухуровневая подсистема хранения (QSPI-NAND для прошивок и «чёрного ящика», SD по SDIO для массивной телеметрии) поддерживает приоритизацию журналирования и энерго-адаптивные режимы записи.

Предложенный разделённый контур „Земля ↔ Борт “снижает сложность верификации, повышает управляемость эксплуатацией и обеспечивает энергобезопасность при жёстких ресурсных ограничениях (Nieto-Peroy et al., 2019; Finance et al., 2021; Dos Santos et al., 2021).

Материалы и методы

БКУ наноспутника проектировался как интегрированная система сбора и предобработки телеметрии, управления энергетикой, оценки и управления ориентацией (ADCS/AOCS), а также исполнения команд и расписаний. Для аппаратов форматов 1U–3U основным конструкторским ограничением выступают ресурсы мощности и тепла, поэтому архитектурные решения выбирались с приоритетом минимизации потребления, предсказуемой отказоустойчивости и проверяемости алгоритмов при использовании компонентной базы класса COTS. В рамках обзора существующих подходов были рассмотрены три доминирующих траектории развития OBC/OBDH: MCU-ориентированные решения на малопотребляющих микроконтроллерах, гибриды CPU+FPGA с реконфигурируемой логикой и, наконец, системы на базе MPU/SoC (Linux-класс). Практика показывает, что для 1U–3U наилучший баланс «производительность/Вт/верифицируемость» обеспечивают контроллерные решения (STM32, MSP430 и аналоги), тогда как гибриды CPU+FPGA оправданы при повышенных требованиях к онборд-обработке и радиационному смягчению, а Linux-класс целесообразен в конфигурациях 6U+ с насыщенными полезными нагрузками (Villela et al., 2019; de Melo et al., 2020; Ali et al., 2014; Sabogal et al., 2020).

Ключевым методологическим принципом стала совместная проработка энергетической подсистемы и БКУ. Это означает приоритизацию критических журналов («чёрного ящика»), аппаратные ограничения токов и «мягкий пуск» тяжёлых линий, а также энерго-обусловленное планирование включений радиолинии и полезной нагрузки. Решения обосновывались на основе системного профилирования потребления и микробенчмаркинга целевого микроконтроллера под реальные режимы ADCS и телеметрии (de Melo et al., 2020; Ali et al., 2014; Sabogal et al., 2020). Для повышения живучести в составе БКУ предусмотрены внешние сторожевые таймеры, порогово-гистерезисные защиты, «безопасные» конфигурации по умолчанию и строго заданная последовательность старта; на уровне ПО применяются автоматизированные стресс- и fuzz-тесты полётного ПО (подтвердившие эффективность, в частности, в проектах SUCHAI), а также стендовая характеристика контуров OBDH под типовые отказные сценарии (Gutierrez et al., 2021; Arseno et al., 2019).

Целевая архитектура реализована вокруг микроконтроллера с насыщенным набором интерфейсов QSPI/SDIO/USB-C/CAN/UART/SPI. Сенсорный контур включает инерциальный модуль (акселерометр и гироскоп), магнитометр, барометр и GPS-приёмник; по каждой оси предусмотрены каналы управления исполнительными механизмами, а также световая индикация

технологических состояний. Подсистема памяти двуслойная: QSPI-FLASH/NAND отведена под прошивки и высоконадёжное журналирование («чёрный ящик»), тогда как карта SD, подключённая по 4-битному SDIO, используется для массивной телеметрии и «сырых» данных. Тактовая база и RTC формируются внешними резонаторами с короткой трассировкой и охранными земляными кольцами. На уровне алгоритмов бортовая часть строго детерминирована: реализованы порогово-гистерезисные защиты, оценка степени заряженности аккумулятора (SoC) на расширенном фильтре Калмана и исполнитель расписаний. Вычислительно затратные задачи - прогноз энергетического баланса и оптимизация расписаний с учётом освещённости и окон связи - вынесены в наземный контур, что упрощает доказательство корректности и снижает риски перегрузки по мощности (Villela et al., 2019; de Melo et al., 2020; Ali et al., 2014; Sabogal et al., 2020).

Обобщённая архитектура БКУ с разнесением интерфейсных доменов, двухуровневой памятью (QSPI-NAND/SD по SDIO), тактовой базой (HSE/LSE) и сенсорным контуром (IMU, магнитометр, барометр, GPS-приёмник) приведена на рисунке 1.

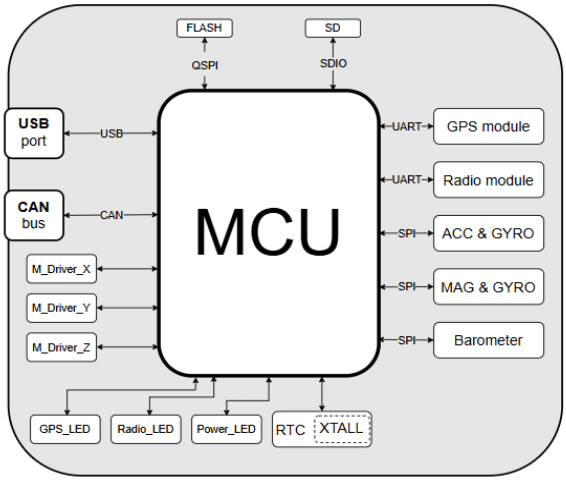


Рис. 1. Архитектура бортового комплекса управления

Контур ориентации построен на сочетании детерминированных оценивателей (TRIAD/QUEST) с рекурсивным уточнением (MEKF/EKF). Такая комбинация обеспечивает требуемую точность при умеренной вычислительной сложности и хорошо согласуется с опытом орбитальной эксплуатации на малых КА (например, UVSQ-SAT и последующие исследования). Для снижения дрейфа гироскопа реализована кросс-проверка показаний IMU и магнитометра; барометр применяется в наземных калибровках и для контроля герметичности при стендовых испытаниях (Takátsy et al., 2022). Последовательность запуска задана как конечный автомат: после стабилизации питания и получения PWR_

GOOD формируется шина 3,3 В и стартует MCU при отключённых «тяжёлых» линиях; далее выполняются самотест и считывание телеметрии, инициализация датчиков, после чего энергоёмкие подсистемы разрешаются условно - по SoC, освещённости, термостанам и прогнозу энергобаланса. Любая аномалия ведёт к немедленному снятию тяжёлых линий и переходу в Safe Mode (Villela et al., 2019; Takátsy et al., 2022).

Конструкторско-технологические решения направлены на обеспечение ЭМС и предсказуемости параметров. Печатная плата выполнена на 4-6 слоях с контролируемыми импедансами по высокоскоростным шинам SDIO/QSPI/USB/CAN и разделением аналоговой и цифровой «земель». В чувствительных цепях применены LC/П-фильтры, на внешних портах - ESD-защита (TVS), на силовых линиях - шунты и датчики тока. Для HSE/LSE обеспечены минимальная длина трасс и локальная развязка (0,1–1,0 мкФ) непосредственно у корпусов. Технология монтажа предполагает использование низколетучих флюсов с обязательной отмывкой и сушкой; высокие компоненты фиксируются (staking), при необходимости наносится конформное покрытие, а кабельные жгуты и разъёмы дополнительно крепятся для повышения вибростойкости (Ali et al., 2014).

Методика верификации включает последовательные функциональные проверки «без АКБ» (с ограничением входного тока) и «с АКБ» (циклы заряд/разряд, UV-lock), климато-механические испытания (термо- и виброциклы), а также кампанию HIL с моделированием динамики ADCS. Дополнительно выполняется профилирование энергопотребления по режимам (ожидание, навигация, приём/передача, работа полезной нагрузки, комбинированные сценарии) и автоматизированный fuzz-тестинг каналов команд/телеметрии и драйверов. Для ADCS планируются количественные метрики - RMS-ошибка ориентации, спектральные характеристики шумов и задержки вычислительного конвейера; для памяти - устойчивость к «brown-out», скорость и надёжность записи в NAND/SD и сохранность «чёрного ящика» в отказных сценариях (Gutierrez et al., 2021; Arseno et al., 2019; Takátsy et al., 2022). Такой план испытаний обеспечивает воспроизводимую оценку функциональной полноты и устойчивости предложенной архитектуры в условиях, близких к эксплуатационным.

В завершение методической части фиксируем принятые допущения по энергобалансу (рабочие шины 3,3 В и 5 В; длительность витка 90 мин с освещённостью ~60 мин и тенью ~30 мин; доли активности подсистем - согласно стендовому профилю). Сводные расчётные значения токов/мощностей по режимам, а также оценка средней мощности и энергии на виток приведены в таблице 1.

Результаты

В настоящей главе представлены ключевые результаты разработки и стендовой отработки БКУ наноспутника формата CubeSat.

Таблица 1. Энергобюджет по режимам (пример для 1U–3U, 3.3 В/5 В шина)

Режим	$I_{3,3В}$, мА	$P_{3,3В}$, Вт	$I_{5В}$, мА	$P_{5В}$, Вт	Мощность, Вт	Доля времени (день/тень)	Средний вклад, Вт
Базовый «Ожидание»	120	0,396	40	0,200	0,596	100 % / 100 %	0,596
ADCS (добавка)	+80	+0,264	+0	+0,000	+0,264	30 % / 20 %	0,069
Радио RX (добавка)	+0	+0,000	+100	+0,500	+0,500	10 % / 6 %	0,066
Радио TX (добавка)	+0	+0,000	+300	+1,500	+1,500	5 % / 0 %	0,050
Полезная нагрузка (добавка)	+0	+0,000	+400	+2,000	+2,000	15 % / 0 %	0,200
Safe Mode (альт. режим)	50	0,165	0	0,000	0,165	-	-
Средняя мощность за виток	-	-	-	-	-	-	≈0,98–1,00
Энергия за виток (90 мин = 1,5 ч)	-	-	-	-	-	-	≈1,47 Вт·ч

Для фокусировки на наиболее информативных аспектах оставлены шесть базовых иллюстраций, отражающих системную интеграцию, сенсорный контур, подсистему хранения, ядро управления, механизм живучести и реализацию печатной платы. Интегральные оценки мощности и энергии по режимам представлены в таблице 1, что позволяет сопоставлять стендовые профили потребления с ограничениями энергетического баланса на виток.

На рисунке 2 приведена укрупнённая структурная схема БКУ с привязкой интерфейсов и каналов питания. Центральным элементом является микроконтроллер (MCU), к которому по шинам QSPI и SDIO подключены энергетически разнесённые носители долговременного хранения: QSPI-FLASH (прошивки, «чёрный ящик») и SD (bulk-телеметрия). Сенсорный контур (IMU, магнитометр, барометр, GPS-приёмник) и коммуникационные устройства (резервная радиолиния, межмодульная шина) подключены по UART/SPI/CAN с соблюдением требований электромагнитной совместимости и ESD-защиты. С учётом допущений таблице 1 средняя мощность за виток составила ~1.0 Вт, при этом доминирующий вклад формирует полезная нагрузка (≈0.20 Вт), далее - ADCS и приёмный тракт (по ≈0.07 Вт каждый), тогда как передача занимает ≈0.05 Вт благодаря малой доле активности. Такая декомпозиция уменьшает сцепление критических и некритических цепей и упрощает верификацию отказных сценариев.

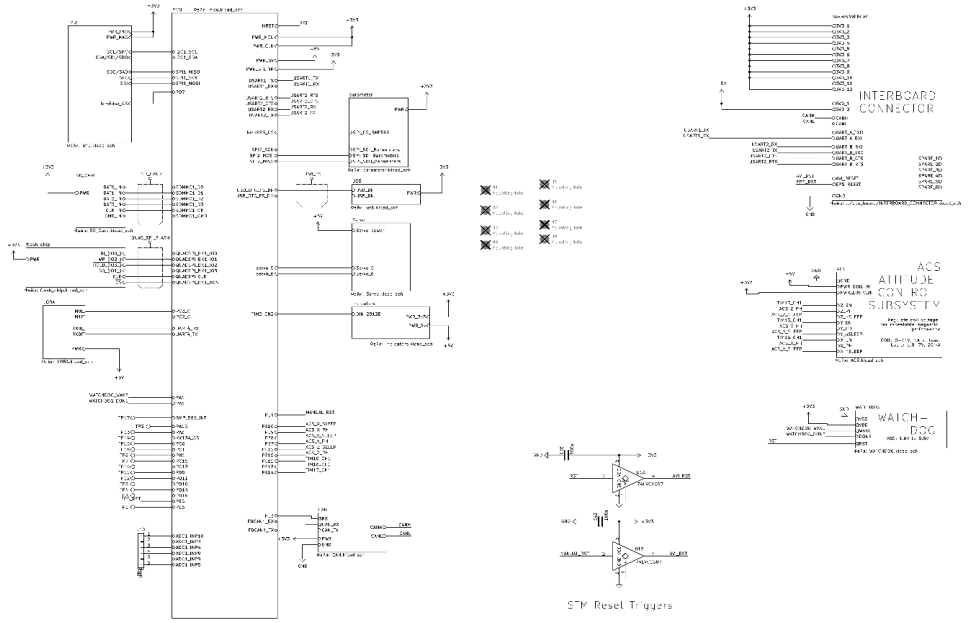


Рис.2. Общая структурная диаграмма подключения функциональных блоков БКУ

Типовое включение LSM6DSV32 (акселерометр/гироскоп) и LSM303AGR (магнитометр) показано на рисунке 3. Линии SPI/I²C оснащены подтягивающими резисторами и RC-фильтрами; питание чувствительных датчиков выполнено от малошумящего LDO с локальной развязкой 0,1–1,0 мкФ у корпусов. Обеспечено физическое разнесение магнитометра относительно токонесущих трасс и силовых ключей для минимизации магнитных наводок. В ходе стендовых испытаний подтверждено соответствие спектральных характеристик шумов паспортным значениям при выбранных режимах дискретизации, а также корректность кросс-коррекции дрейфа гироскопа по данным магнитометра.

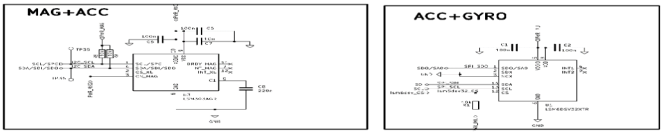


Рис. 3. Типовая схема подключения IMU LSM6DSV32 и магнитометра LSM303AGR

Интегральная схема подсистемы хранения представлена на рисунке 4: 4-битный SDIO → SD-карта с ESD-защитой сигнальных линий и согласованием импедансов, а также QSPI → SLC-NAND с линиями WP#/HOLD# и отдельной фильтрацией питания. Программная логика использует двухуровневый буфер ОЗУ и приоритизацию сбросов: критические события немедленно записываются в NAND, тогда как телеметрия агрегируется на SD. Такой режим обеспечил сохранность аварийных дампов при имитации дефицита энергии за счёт снижения частоты операций записи на SD.

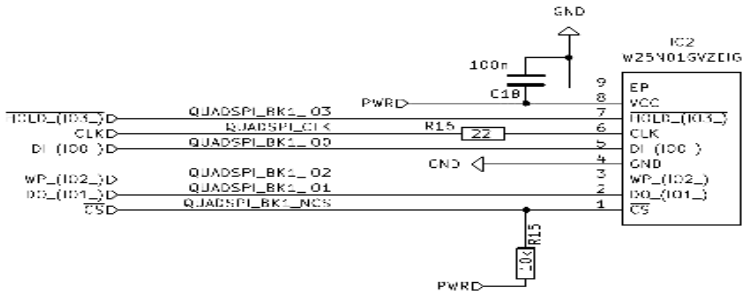


Рис. 4. Типовая схема подключения NAND Флеш памяти по интерфейсу qspi

На рисунке 5 демонстрируется включение STM32H743VIT с внешним тактированием (HSE 25 МГц, LSE 32,768 кГц) и интерфейсом SWD. Трассировка резонаторов выполнена по короткому пути с охранными земляными кольцами; для HSE соблюдены требования по нагрузочным конденсаторам. Подтверждена надёжная инициализация контроллера в холодном/тёплом старте, и корректная работа FPU при целевой частоте до 400 МГц.

Механизм живучести реализован на базе внешнего сторожевого таймера TPL5010 (на рисунке 6). Окно наблюдения задаётся резистивной матрицей; «пинок» формируется по линии WAKE, аварийный сброс - по RST. На стенде таймер устойчиво детектировал зависания пользовательского потока и инициировал восстановление работоспособности; собственное потребление узла соответствовало нанопотреблению, что критично для длительного пребывания в Safe Mode.

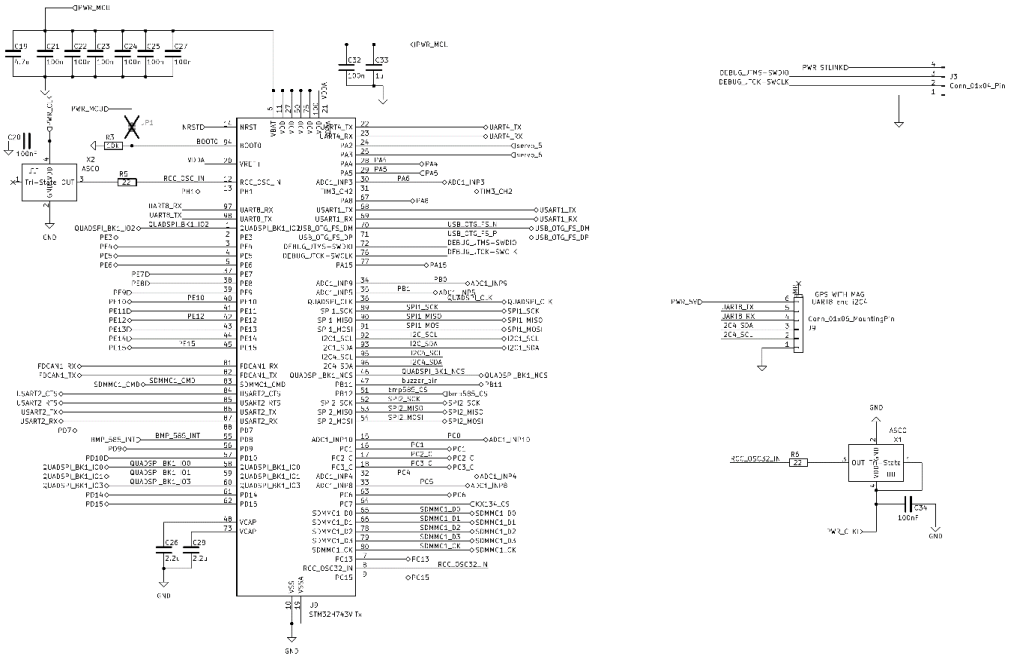


Рис. 5. Типовая схема подключения MCU STM32H743VIT с источниками внешнего тактирования (HSE, LSE)

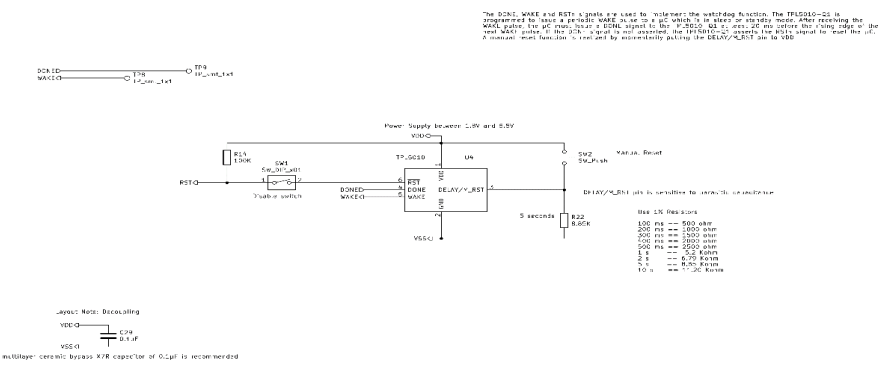
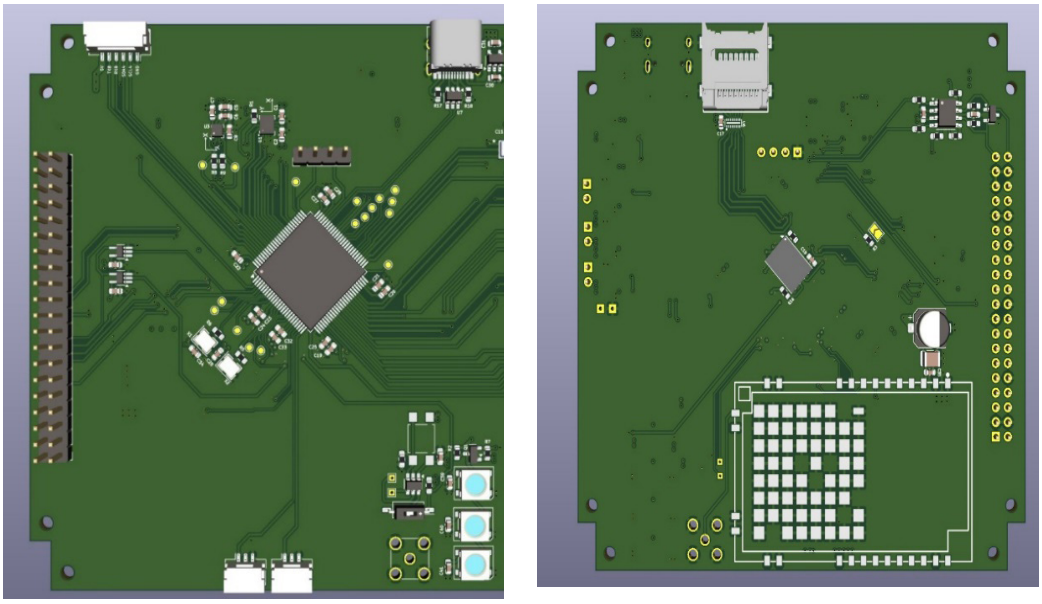


Рис. 6. Типовая схема включения внешнего сторожевого таймера TPL5010

Общий вид печатной платы представлен на рисунке 7. Плата выполнена на 4–6 слоях с контролируемыми импедансами по высокоскоростным шинам

(SDIO/QSPI/USB/CAN). Обеспечены разделение аналоговой и цифровой «земель», короткие тракты HSE/LSE, а также экранирование областей датчиков от доменов с высокими dV/dt. Размещение RTC и индикаторов выполнено с учётом минимизации паразитных наводок на чувствительные цепи.



(а) (б)

Рис. 7. Общий вид печатной платы БКУ: (а) верхняя сторона; (б) нижняя сторона

Проведённая функциональная верификация подтвердила корректность ключевых сценариев работы. Последовательность запуска реализуется строго по конечному автомату: после стабилизации входного питания и получения PWR_GOOD подаётся шина 3,3 В, стартует MCU при отключённых «тяжёлых» линиях, выполняются самотест и считывание телеметрии, далее инициализируются датчики; разрешение энергоёмких подсистем происходит условно - по состоянию аккумулятора (SoC), освещённости, температурным ограничениям и прогнозу энергетического баланса. Канал обмена с сенсорами (IMU, магнитометр, барометр) показал стабильность: спектральные характеристики шумов соответствуют паспортным данным, встроенные цифровые фильтры IMU функционируют корректно. Подсистема хранения данных продемонстрировала требуемую устойчивость: критические события надёжно фиксируются в QSPI-NAND при снижении частоты сбросов на SD в условиях энергетического дефицита - «чёрный ящик» не теряется. Внешний сторожевой таймер обеспечил детектирование зависаний и инициировал восстановление при собственном нано-уровне потребления, совместимом с длительным Safe Mode.

Суммарно результаты указывают на достаточную функциональную

полноту предлагаемой архитектуры для аппаратов формата 1U–3U при жёстких энергетических ограничениях. Сенсорный контур на базе компонентов COTS обеспечивает устойчивую работу 9-DoF-оценки ориентации с возможностью рекурсивного уточнения, а приоритизация журналирования и разделение критичных и массовых потоков данных повышают устойчивость к отказам и управляемость восстановлением. Реализованная печатная плата подтверждает технологическую осуществимость схемных решений и соответствие требованиям EMC/ESD, формируя основу для расширенной программы климатических и вибродинамических испытаний на следующем этапе.

Обсуждение

В дальнейшей работе планируется завершить комплексную прошивку и интеграцию интеллектуального контура управления. На низком уровне будут отлажены загрузчик с резервным образом в QSPI-NAND и безопасное обновление (DFU по USB-C, резервный канал по LoRa), реализованы драйверы HAL для SDIO/QSPI/CAN/SPI/I²C и подсистема синхронизации (HSE/LSE), а также внедрён энерго-ориентированный планировщик (порог/гистерезис + оценка SoC на EKF) и штатные процедуры восстановления (внешний watchdog TPL5010, приоритизированное журналирование «чёрного ящика»). На уровне алгоритмов предусмотрены интеграция ADCS (TRIAD/QUEST + MEKF) с калибровкой датчиков и аппаратными HIL-испытаниями, а также верификация устойчивости в термо-виброциклах и при инъекциях отказов памяти. Контур ИИ предполагается развернуть преимущественно на наземной станции: прогноз освещённости и энергетического баланса, адаптивная оптимизация расписаний (стохастический поиск в сочетании с ограниченным RL), раннее обнаружение аномалий телеметрии (self-supervised/autoencoder) и выдача рекомендаций по настройке порогов с последующей телеметрически обоснованной валидацией; на борту допустима лишь лёгкая инференс-прослойка (компрессия/квантизация моделей) в некритическом контуре. Дополнительно запланированы меры кибербезопасности (подписанные прошивки, контроль целостности), расширенные испытания долговечности носителей (SD/NAND) и выпуск пакета воспроизводимости (скрипты сборки, профили нагрузок, трассируемые протоколы испытаний), по результатам которых будет сформирован релиз-кандидат программно-аппаратной платформы.

Заключение

Представленная работа последовательно обосновывает и реализует архитектуру бортового комплекса управления для наноспутников формата CubeSat с акцентом на энергетическую безопасность, верифицируемость и отказоустойчивость. Компоновка на базе MCU STM32H743 в сочетании с внешним сторожевым таймером, разнесённая по приоритетам подсистема хранения (QSPI-NAND для критических журналов и SD по SDIO для массивной телеметрии), а также сенсорный контур (IMU, магнитометр, барометр) продемонстрировали состоятельность на стенде: соблюдение автомата

запуска, стабильный обмен с датчиками, сохранность «чёрного ящика» при энергетическом дефиците и надёжное восстановление после сбоев. Выбранная алгоритмическая схема - детерминированные оцениватели ориентации (TRIAD/QUEST) с рекурсивным уточнением (MEKF/EKF) - обеспечивает требуемую точность при ограниченных ресурсах, а вынесение вычислительно затратного планирования на наземную станцию снижает риски и упрощает доказательство корректности критических функций.

Ограничения текущего этапа связаны с отсутствием длительных ресурсных испытаний (температурные и вибродинамические циклы, радиационная стойкость) и полноценной HIL-кампании с исполнительными механизмами, а также с необходимостью накопления расширенной статистики для численной оценки энергобаланса и шумовых характеристик датчиков. Запланированные шаги - завершение прошивки с безопасным обновлением и контролем целостности, интеграция наземного ИИ-контура (прогноз энергии, адаптивные расписания, раннее выявление аномалий) при сохранении жёсткой детерминированности бортовой логики, а также расширенные испытания долговечности носителей (SD/NAND) и устойчивости к отказам - доведут платформу до уровня релиз-кандидата для миссий 1U–3U с жёсткими ограничениями по массе, энергии и допустимому риску.

REFERENCES

Ali, A., Mughal, M.R., Ali, H. & Reyneri, L. (2014). Innovative power management, attitude determination and control tile for CubeSat standard NanoSatellites. Acta Astronautica. — 96. — Pp. 116–127. https://doi.org/10.1016/j.actaastro.2013.11.013

Belokonov, I., Timbai, I. & Nikolaev, P. (2018). Analysis and synthesis of motion of aerodynamically stabilized nanosatellites of the CubeSat design. Gyroscopy and Navigation. — 9. — Pp. 287–300. https://doi.org/10.1134/S2075108718040028

de Melo, A.C.C.P., Café, D.C. & Borges R.A. (2020). Assessing power efficiency and performance in nanosatellite onboard computer for control applications. IEEE Journal on Miniaturization for Air and Space Systems. — 1(2). — Pp. 110–116. https://doi.org/10.1109/JMASS.2020.3009835

Dos Santos, G.H., Seman, L.O., Bezerra, E.A., Leithardt, V.R.Q., Mendes, A.S. & Stefenon, S.F. (2021). Static attitude determination using convolutional neural networks. Sensors. — 21(19). — P. 6419.

Finance, A., Dufour, C., Boutéraon, T., Sarkissian, A., Mangin, A., Keckhut, P. & Meftah, M. (2021). In-orbit attitude determination of the UVSQ-SAT CubeSat using TRIAD and MEKF methods. Sensors. — 21(21). — P. 7361. https://doi.org/10.3390/s21217361

Gutierrez, T., Bergel, A., Gonzalez, C.E., Rojas, C.J., & Diaz, M.A. (2021). Systematic Fuzz Testing Techniques on a Nanosatellite Flight Software for Agile Mission Development. IEEE Access. — 9. — Pp. 114008–114021. https://doi.org/10.1109/ACCESS.2021.3104283

Liddle, J.D., Holt, A.P., Jason, S.J. et al. (2020). Space science with CubeSats and nanosatellites. Nature Astronomy. — 4. — Pp. 1026–1030. https://doi.org/10.1038/s41550-020-01247-2

Nieto-Peroy, C. & Emami, M.R. (2019). CubeSat mission: From design to operation. Applied Sciences. — 9(15). — P. 3110. https://doi.org/10.3390/app9153110

Poghosyan, A., & Golkar, A. (2017). CubeSat evolution: Analyzing CubeSat capabilities for conducting science missions. Progress in Aerospace Sciences. — 88. — Pp. 59–83. https://doi.org/10.1016/j.paerosci.2016.11.002

Rigo, C.A., Seman, L.O., Camponogara, E., Morsch Filho, E., Bezerra, E.A. & Munari, P. (2022). A branch-and-price algorithm for nanosatellite task scheduling to improve mission quality-of-service. European Journal of Operational Research. — 303(1). — Pp. 168–183.

Sabogal, S., George, A. & Wilson, C. (2020). Reconfigurable framework for environmentally adaptive resilience in hybrid space systems. ACM Transactions on Reconfigurable Technology and Systems. — 13(3). — Pp. 1–32. https://doi.org/10.1145/339838

Takátsy, J., Bozóki, T., Dálya, G. et al. (2022). Attitude determination for nano-satellites — II. Dead reckoning with a multiplicative extended Kalman filter. Experimental Astronomy. — 53. — Pp. 209–223. https://doi.org/10.1007/s10686-021-09818-5

Villela, T., Costa, C.A., Brandão, A.M., Bueno, F.T. & Leonardi, R. (2019). Towards the thousandth CubeSat: A statistical overview. International Journal of Aerospace Engineering. — 2019. — Pp. 1–13. https://doi.org/10.1155/2019/5063145

Arseno, D., Edwar E., Harfian, A.R. & Salsabila, J.N. (2019). Characterization of On-Board Data Handling (OBDH) Subsystem. In 2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA). — Bali, Indonesia. — Pp.. 223–227. https://doi.org/10.1109/TSSA48701.2019.8985466

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 153–173
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.009
UDC 004.021

MATHEMATICAL FILTERING IN CALL HISTORY FORENSICS MODULE

T. Grigoryev[1], P. Tazhibayeva[1*], A. Imanberdi[2], R. Lisnevskyi[1]
[1,]Astana IT University, Astana, Kazakhstan;
[2]TSARKA GROUP, Astana, Kazakhstan.
E-mail: 242924@astanait.edu.kz

Abulkair Imanberdi — Master of Science, head of the team of "TSARKA" GROUP, Astana, Kazakhstan
https://orcid.org/0009-0005-6144-2392;
Timur Grigoriev — Master's student, BSc, researcher, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0002-1082-5295;
Tazhibayeva Perizat — Master's student, BSc, researcher, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0009-1566-636X;
Lisnevskyi Rostyslav — PhD, associate professor, Cybersecurity School, Astana IT University, Astana, Kazakhstan
https://orcid.org/ 0000-0002-9006-6366.

Abstract. The growing sophistication of technology-enabled crime demands that forensic practitioners move beyond data extraction toward analytical interpretation. This study evaluates four major suites - Cellebrite UFED, Oxygen Forensic Detective, Elcomsoft iOS Forensic Toolkit, and MOBILedit Pro - focusing on their ability to parse call logs from modern Android and iOS devices. While reliable in acquisition, these tools show recurring limitations: incomplete reconstruction of fragmented SQLite tables, weak temporal visualization, and poor cross-device correlation. To address these issues, we propose a lightweight Call History Analysis Module that integrates into existing workflows. It unifies heterogeneous call databases into a chronological ledger, applies mathematical filters to detect patterns or anomalies, and visualizes communication trends via heatmaps, duration ridgelines, and ego-network graphs. A built-in similarity search engine identifies shared contacts and parallel timelines. Pilot testing confirms improved accuracy, reduced triage time, and enhanced contextual insight, extending digital forensic capabilities at minimal operational cost.

Keywords: forensics, mobile, calls, analysis, filtering, visualization, evidence

For citation: T. Grigoryev, P. Tazhibayeva, A. Imanberdi, R. Lisnevskyi. Mathematical Filtering in Call History Forensics Module//International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 153–173.

(In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.009.

Conflict of interest: The authors declare that there is no conflict of interest.

ҚОҢЫРАУЛАР ТАРИХЫНДАҒЫ СОТ САРАПТАМАСЫ МОДУЛІНДЕГІ МАТЕМАТИКАЛЫҚ СҮЗУ

Т. Григорьев[1], П. Тажибаева[1], А. Иманберді[2], Р. Лисневский[1]
[1]Astana IT University, Астана, Қазақстан;
[2]TSARKA GROUP, Астана, Қазақстан.
E-mail: 242924@astanait.edu.kz

Иманберді Әбілқайыр — Ғылым Магистрі, «ЦАРКА» ТОБЫ ұжымының жетекшісі
https://orcid.org/0009-0005-6144-2392;
Григорьев Тимур — магистрант, бакалавр, Астана IT Университетінің ғылыми қызметкері
https://orcid.org/0009-0002-1082-5295;
Перизат Тажибаева — магистрант, бакалавр, Astana IT Университетінің ғылыми қызметкері
https://orcid.org/0009-0009-1566-636x;
Лисневский Ростислав — PhD, Астана IT университетінің киберқауіпсіздік мектебінің доценті
https://orcid.org/ 0000-0002-9006-6366.

Аннотация. Технологияны қолдана отырып жасалған қылмыстардың өсіп келе жатқан күрделілігі сот-медициналық сарапшылардан деректерді жинау шеңберінен шығып, аналитикалық интерпретацияға көшуді талап етеді. Бұл зерттеу төрт негізгі жиынтықты бағалайды - Cellebrite UFED, Oxygen Forensic Detective, Elcomsoft Ios Forensic Toolkit және MOBILedit Pro - олардың заманауи Android және iOS құрылғыларындағы қоңыраулар журналдарын талдау қабілетіне бағытталған. Сатып алу кезінде сенімді болғанымен, бұл құралдар қайталанатын шектеулерді көрсетеді: Фрагменттелген SQLite кестелерінің толық емес қайта құрылуы, әлсіз уақытша визуализация және құрылғылар арасындағы нашар корреляция. Осы мәселелерді шешу үшін біз Бар Жұмыс процестеріне біріктірілген Қоңыраулар Тарихын Талдаудың жеңіл Модулін ұсынамыз. Ол қоңыраулардың гетерогенді дерекқорларын хронологиялық кітапқа біріктіреді, заңдылықтарды немесе ауытқуларды анықтау үшін математикалық сүзгілерді қолданады және жылу карталары, ұзақтық сызықтары және эго-желілік графиктер арқылы байланыс тенденцияларын визуализациялайды. Кірістірілген ұқсастық іздеу жүйесі ортақ контактілерді және параллель уақыт кестелерін анықтайды. Пилоттық тестілеу дәлдіктің жақсарғанын, сұрыптау уақытының қысқарғанын және контекстік түсініктің жақсарғанын растайды, бұл цифрлық сот сараптамасының мүмкіндіктерін ең аз

операциялық шығындармен кеңейтеді.

Түйін сөздер: Криминалистика, Ұялы Байланыс, Қоңыраулар, Талдау, Сүзу, Визуализация, Дәлелдемелер

Дәйексөздер үшін: Ә. Иманберді, Т. Григорьев, П. Тажибаева, Р. Лисневский. Қоңыраулар Тарихындағы Сот Сараптамасы Модуліндегі Математикалық Сүзу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 153–173 бет. (Ағыл). https:// doi. org/10.54309/IJICT.2025.24.4.009.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

МАТЕМАТИЧЕСКАЯ ФИЛЬТРАЦИЯ В МОДУЛЕ СУДЕБНОЙ ЭКСПЕРТИЗЫ ИСТОРИИ ЗВОНКОВ

Т. Григорьев[1], П. Тажибаева[1], А. Иманберди[2], Р. Лисневский[1]

[1]Астана ИТ университет, Астана, Казахстан;
[2]TSARKA GROUP, Астана, Казахстан.
Email: 242924@astanait.edu.kz

Иманберди Абулкаир — магистр наук, руководитель команды группы компаний «ЦАРКА»
https://orcid.org/0009-0005-6144-2392;
Тимур Григорьев — магистрант, бакалавр, исследователь, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0002-1082-5295;
Тажибаева Перизат — магистрант, бакалавр, исследователь, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0009-1566-636x;
Ростислав Лисневский — кандидат технических наук, доцент, Школа кибербезопасности, Астана ИТ университет, Астана, Казахстанhttps://orcid.org/ 0000-0002-9006-6366.

Аннотация. Растущая сложность преступлений, совершаемых с использованием современных технологий, требует от судебно-медицинских экспертов перехода от извлечения данных к аналитической интерпретации. В этом исследовании оцениваются четыре основных пакета - Cellebrite UFED, Oxygen Forensic Detective, Elcomsoft iOS Forensic Toolkit и MOBILedit Pro - с акцентом на их способность анализировать журналы вызовов с современных устройств Android и iOS. Несмотря на надежность сбора данных, эти инструменты имеют ряд недостатков: неполное восстановление фрагментированных таблиц SQLite, слабая временная визуализация и плохая корреляция между устройствами. Для решения этих проблем мы предлагаем легкий модуль анализа истории звонков, который интегрируется в существующие рабочие процессы. Он объединяет

разнородные базы данных о звонках в хронологический реестр, применяет математические фильтры для выявления закономерностей или аномалий и визуализирует тенденции обмена сообщениями с помощью тепловых карт, временных линий и графиков эго-сети. Встроенная система поиска сходства идентифицирует общие контакты и параллельные временные линии. Пилотное тестирование подтверждает повышение точности, сокращение времени сортировки и улучшение понимания контекста, расширяя возможности цифровой криминалистики при минимальных эксплуатационных затратах.

Ключевые слова: криминалистика, телефон, звонки, анализ, фильтрация, визуализация, доказательства

Для цитирования: Т. Григорьев, П. Тажибаева, А. Иманберди, Р. Лисневский. Математическая фильтрация в модуле судебной экспертизы истории звонков//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 23. Стр. 153–173. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.009.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Due to the exponential rise in cybercrime, digital forensics plays a critical role in supporting legal investigations and ensuring evidentiary integrity (Batra: 2020: 42–46). This involves analyzing various devices, including computers, tablets, SIM cards, and, most importantly, mobile phones, which have become integral to daily life due to their widespread use and large storage capacities (Alatawi: 2020: 1–6). Digital forensics plays a critical role in modern law enforcement, cybersecurity, and legal investigations and offers the technical framework to ensure the integrity and reliability of digital evidence (Mahmoud: 2023: 1–6).

Data extraction and analysis are the two primary stages of digital forensic investigations, and global trends in these areas should be reviewed (Shimmi: 2020: 1–7). It is important to take into consideration that the second stage was developed only after the development of modern technologies for accounting and registration of new digital information about a person and his environment, which are generated in automatic, semi-automatic and manual mode and in unlimited quantities. As a result, nowadays the stage of forensic analysis is equally important with the extraction stage (Khalaf: 2019: 1–5).

For example, artificial intelligence enhances data analysis, blockchain ensures evidence authenticity, and cloud solutions facilitate the storage and processing of large amounts of information (Tyagi: 2022: 1–6).

Data analysis technologies, particularly those leveraging artificial intelligence, are increasingly critical and widely adopted (Rizvi: 2022: 110362–110384). For instance: AI helps automate routine tasks such as extracting and classifying data from various sources, can perform semantic analysis, especially analyzing text messages and emails to identify potential malicious content or threat (Waluyo: 2022: 95–100). In addition, machine learning models can be used to predict likely locations and types of crimes based on historical data (Punjabi: 2022: 1–5).

There are a number of important fields that is needed to be taken into consideration for digital forensics analysis:

1. Artificial Intelligence (AI) and Machine Learning (ML)
2. Cloud technologies
3. Internet of Things (IoT)
4. Blockchain

Artificial Intelligence (AI) and Machine Learning (ML) are revolutionizing digital forensics by automating complex tasks and increasing the accuracy of investigations (Rizvi: 2022: 110362–110384).

Cloud technologies are also important, because more and more organizations move their data and application to the cloud, cloud forensics is becoming an important area of attention. It presents unique challenges for criminologists, such as data variability, multi-user architecture and jurisdiction issues (Fernandes: 2020: 422–427).

Internet of Things forensics involves extracting and analyzing data from interconnected devices such as smart home systems, wearable technologies, and industrial control systems. These devices generate huge amounts of data, and forensic investigators must develop specialized methods to process the variety and volume of information (Fernandes: 2020: 422–427).

Blockchain Technology is known for its secure and immutable registry and gaining momentum in digital forensics. It is relevant for investigation of crimes related to cryptocurrency, such as fraud, money laundering, corruption and attacks using ransomware. Furthermore, blockchain is used to ensure the integrity and authenticity of digital evidence (Hou: 2020: 1–15).

New technologies, such as tamper-proof storage and advanced hashing algorithms, are being developed to enhance the security and integrity of digital evidence. These include tamper- proof data storage solutions, blockchain-based proof chains, and advanced hashing algorithms to verify data integrity (Alatawi: 2020: 1–6).

Basic call, text message, and contact data are stored on mobile devices, but smart phones retain considerably more valuable data, including chat logs, email, personal information, and browser history.

The article (Mallidi: 2016: 1–12) highlights the need for forensic tools tailored to smartphones for data examination and recovery. Analyzing call history is crucial in mobile forensics, as call logs reveal communication patterns, connections, and time- lines. Overlooking key insights in call history analysis can impact the entire forensic process.

Methods and materials

Modern software products for mobile forensics allow experts to receive and analyze data from mobile devices, which plays a key role in digital investigations. Among the many solutions, several popular tools stand out, which have their own characteristics and fields of application (Cristian: 2020: 1–7).

A. Cellebrite UFED

Cellebrite has developed a series of products called the Universal Forensic Extraction Device (UFED), which are used for mobile forensic examinations (Tara: 2021: 97–104).

Key Features:

• Physical abstraction and decrypting while bypassing the pattern lock

• Set of analysis characteristics

• Advanced searching and filtering to quickly locate critical data (keywords, dates, specific data types)

• Timeline Analysis, powerful reporting capabilities

• Supports more than 30,000 device models

• Ability to work with deleted data and backups

B. Oxygen Forensic Tool

Oxygen Forensic Tool provides deep data analysis, that includes application data and records (Oxygen Forensics: 2024).

Key Features:

• Extract data from over 31.000 devices

• Data extraction from drones, IoT Devices, Smartwatches, Fitness Apps

• Finds, extracts and decrypts passwords, credentials, system files, and user data from Windows, macOS, or Linux

• Optical Character Recognition (OCR)

• Statistic widgets, Social Graph interface

• Timeline analytics

• Identification of frequently visited places, specify time periods, common locations and animate travels

• Face categorization and analytics (gender, race, age)

C. Elcomsoft IOS Forensic Toolkit

Elcomsoft IOS Forensic Toolkit is a program for extracting data from iOS devices, including physical extraction from locked phones, as well as working with iTunes and iCloud backups (Lahaie: 2015: 1–50).

Key Features:

• Passcode unlock: Brute-forces 4-digit and 6-digit screen lock passcodes via DFU exploit.

• Full file system extraction and keychain decryption for many devices running iOS 12 through 16.5.

• Media extraction, shared files, and advanced logical acquisition for iOS devices running versions that the extraction agent does not support.

• Has the ability to extract saved files and crash/diagnostics logs from many apps. Extract documents locally stored in Adobe Reader and Microsoft Office, access the Mini-KeePass password database, and much more.

D. MOBILedit Pro

MOBILedit Pro is a forensic tool that is designed for phone and cloud extraction, data analyzing and report generating, uses both physical and logical data acquisition methods (Hermawan: 2020: 233–238).

Key Features:

• Security bypassing – acquires physical image even when phone is protected by a password or pattern

• Physical data acquisition (extract physical images and have exact binary clones) and analysis (open image files created by this process)

• Advanced application analysis (adaptive and in-depth methods to ensure retrieving the most data available).

Focus on timeline as a note, a photo, video or a flow of messages

• Walkthrough for applications, photos, accounts, contacts, messages, emails,

calls and organizers

- Malware detection
- Scientific image analysis
- Has photo recognizer and photo matcher

A comparative analysis reveals that, although all four forensic tools excel in specific areas, they also exhibit notable weaknesses. Some of them are great in data extraction others in data analyzing. They all have shortcomings where other tools perform better. One main common problem is a wide range of devices under study, that happens because of constant changes and updates of devices (Delija: 2021: 1231–1235).

Additionally, challenges vary depending on the task, including gaining access to or unlocking devices, searching and filtering data, analyzing specific elements, and generating reports. Furthermore, none of them provide a comprehensive solution for the deep call history analysis.

Each tool has their own strength, but when it comes to detailed examination of call logs they all fail. Even though, they all support general call extraction, none of the tools specifically focus on those points:

1. Comprehensive analysis of call metadata, that includes: frequency of calls between specific users and call duration patterns

2. Identifying call trends through visualization and reporting on call interactions

3. Advanced filtering call related data that includes missed or rejected calls(-can be helpful during forensic investigations)

Recent research emphasizes the necessity of specialized modules for analyzing call history, focusing on mathematical filtering, semantic search, and statistical visualization, which traditional forensic tools still lack (Rzayeva: 2025: 66–74).

According to everything that is written above, it follows that there are no universal methods for solving these problems in the market within the framework of the tasks being solved.

To thoroughly analyze a mobile device, relying on two to three (or even five) programs is often insufficient. The organization that provided some of these tools and test data asked to combine all the useful functions of analyzing and tracking phone calls in one module.

Below is a table of functions and a list of tested applications. Unfortunately, the data used in the tests is personal data that is used in the investigation. According to the law of the Republic of Kazakhstan, these data cannot be published.

Table 1. «Comparison of forensic analysis tools»

Feature	UFED 4PC	MOBILedit Pro	Oxygen forensic
Average call duration	+	-	+
Call types (incoming, outgoing)	+	+	+
Call apps used	+	-	+
Top 10 contacts	+	-	+
Activity periods	-	-	-
Total calls periods	+	+	+

International calls location detection	-	-	-
Kazakhstan city calls detection	-	-	-

Results

The proposed Call History Analysis Module enables investigators to efficiently filter and analyze large datasets of call records using phone number and timestamp conditions. Key features such as phone number filtering, duration statistics and app usage analysis helps to give a comprehensive picture of communication patterns. Furthermore, the module provides insights into important contacts, call durations, and active communication periods, which are visualized through charts. This method not only makes call history analysis easier, but it also improves investigators' capacity to swiftly pinpoint important contacts and communication patterns. In the future, functions will be developed to determine the locations of international calls and cities in Kazakhstan. All tested tools do not have this feature.

A. Call history analyzer

The following functions are demonstrated on dataset $D = \{d_1, d_2, \ldots d_n\}$, where each element d_i has the following attributes:

- $d_{i.number}$: The phone number involved in the call..
- $d_{i.timestamp}$:The timestamp of the call in ISO format.
- $d_{i.type}$:, which can be either "incoming" or "outgoing".
- $d_{i.duration}$: The duration of the call (in seconds).
- $d_{i.app}$: The application used for the call.

Let:

- N be the given phone number filter (optional).
- t_s be the given start time (optional).
- t_e be the given end time (optional).
- $ISO(t)$ be a function that converts a timestamp string t to an ISO datetime object for comparison.

The filtered dataset F can be expressed as:

$$F = \{d_i \epsilon D \mid (N = \emptyset \vee d_i.number = N) \wedge (t_s = \emptyset \vee t_e = \emptyset \vee t_s \leq ISO(d_i.timestamp) \leq t_e)\} \quad (1)$$

Where:

- $N = \emptyset$ means no phone number filter is applied
- $t_s = \emptyset$ or $t_e = \emptyset$ means no time boundary is set.

Explanation

1) Phone Number Filtering:
   - If N is provided, only records where $d_i.number = N$ are included.
   - If N is not provided, all records are included.
2) Time Range Filtering:
   - If both t_s and t_e are provided, the records must satisfy:

- $t_s \leq ISO(d_i.timestamp) \leq t_e$
- If only one boundary is given (either t_s or t_e), the records are filtered accordingly.

1) Total Call Counts: The number of incoming and outgoing calls. The total number of incoming and outgoing calls in the filtered dataset is calculated. To do this, iterate over each call record and increase the increment responsible for each type of call:

Incoming calls: Calls whose type has been marked as ”incoming”.
Outgoing calls: Calls whose type has been marked as ”outgoing”.
Mathematical Representation of get_incoming_outgoing_calls
Let:

- C_{in}represents the total number of incoming calls.
- C_{out} represents the total number of outgoing calls.

The function calculates:

$$C_{in} = \sum_{d_i \in D} 1(d_i.type = "incoming") \quad (2)$$

Formula (2) calculates C_{in} by adding 1 for every call record whose type is “incoming,” giving the total count of inbound calls.

$$C_{out} = \sum_{d_i \in D} 1(d_i.type = "outgoing") \quad (3)$$

Formula (3) determines C_{out} by adding 1 for each record labeled “outgoing,” producing the total number of outbound calls.

where:

$$1(P) = \begin{cases} 1 \ if \ P \ is \ true \\ 0 \ if \ P \ is \ false \end{cases} \quad (4)$$

Formula (4) defines the indicator function $1(P)$, which returns 1 when condition P is true and 0 when it is false.

Explanation:

1) Incoming Calls:

$$C_{in} = \sum_{d_i \in D} 1(d_i.type = \text{incoming}) \quad (5)$$

counts all records where the call type is “incoming”.

Formula (5) sums the indicator $1(d_i.type$ =incoming) over every record $d_i \in D$, yielding the total number of incoming calls C_{in}.

2) Outgoing Calls:

$$C_{out} = \sum_{d_i \in D} 1(d_i.type = "outgoing") \quad (6)$$

counts all records where the call type is "outgoing".

Formula (6) sums the indicator $1(d_i.type$="outgoing") over every record $d_i \in D$, yielding the total number of outgoing calls C_{out}.

The function returns the tuple (C_{in}, C_{out}), representing the total counts of incoming and outgoing calls, respectively.

2) *Call Duration:* Detailed statistics on the duration of calls have been calculated to assess the duration and patterns of communication. The statistics included:

Total number of calls: the total number of analyzed calls.

Average duration: The average duration of calls calculated by summing all call durations and dividing by the total number of calls.

Maximum and minimum durations: The longest and shortest of the observed call durations.

Calculations are performed for all calls in aggregate and separately for incoming and outgoing calls.

Mathematical Representation of "get_call_duration_statistics"

Let:

- C_{in} is the total number of incoming calls.
- C_{out} is the total number of outgoing calls.
- $n = |D|$ is the total number of calls.

The function computes the following statistics:

a) Total Calls Statistics:

Total Amount = n

$$Average\ Duration = \frac{1}{n}\sum_{i=1}^{n} d_i.duration \quad (7)$$

Formula (7) finds the overall average call length by dividing the sum of all call durations by the total number of calls n.

$$Max\ Duration = max\{d_i.duration | d_i \in D\} \quad (8)$$

Formula (8) returns the single longest call duration present in the entire data set D.

$$Min\ Duration = min\{d_i.duration | d_i \in D\} \quad (9)$$

Formula (9) returns the single shortest call duration present in D.

b) Incoming Calls Statistics:

$$Incoming\ Amount = C_{in} \tag{10}$$

Formula (10) simply states that the total number of incoming calls equals C_{in}

$$Average\ Incoming\ Duration = \frac{1}{C_{in}} \sum_{\substack{i=1 \\ d_i.type="incoming"}}^{n} d_i.duration \tag{11}$$

Formula (11) computes the mean duration of incoming calls by averaging only those records whose type is "*incoming*."

$$Max\ Incoming\ Duration = max\{d_i.duration | d_i.type = "incoming"\} \tag{12}$$

Formula (12) selects the maximum duration among all incoming calls.

$$Min\ Incoming\ Duration = min\{d_i.duration | d_i.type = "incoming"\} \tag{13}$$

Formula (13) selects the minimum duration among all incoming calls.

c) Outgoing Calls Statistics:

$$Average\ Outgoing\ Duration = \frac{1}{C_{outgoing}} \sum_{\substack{i=1 \\ d_i.type="outgoing"}}^{n} d_i.duration \tag{14}$$

Formula (14) computes the mean duration of outgoing calls by averaging only those records whose type is "outgoing."

$$Max\ Outgoing\ Duration = max\{d_i.duration | d_i.type = "outgoing"\} \tag{15}$$

Formula (15) selects the maximum duration among all outgoing calls.

$$Min\ Outgoing\ Duration = min\{d_i.duration | d_i.type = "outgoing"\} \tag{16}$$

Formula (16) selects the minimum duration among all outgoing calls.

d) Handling Special Cases:

- If $D = \emptyset$, all statistics are set to 0.
- If $min\{d_i.duration \mid d_i \in D\} = \infty$, it is reset to 0.

3) Application Usage: In some systems, the applications (Telegram, WhatsApp, etc.) used for the call may be present in the call history. In this case, for each application registered in the call data, the number of incoming and outgoing calls made using this application is calculated.

Mathematical Representation of get_call_apps

Let:

- $A = \{a_1, a_2, \ldots, a_m\}$ be the set of unique apps in the dataset.
- calls(a_j ,t) be the number of calls for app $a_j \in$ A and
- type $t \in$ {“incoming”, “outgoing”}.

The function computes the following:

$$calls(a_j, t) = \sum_{\substack{i=1 \\ d_i.app = a_j \\ d_i.type = t}}^{n} 1 \quad (17)$$

Formula (17) counts how many calls of type tt are handled by application a_j

Where:

$$t \in \{"incoming", "outgoing"\} \quad (18)$$

Formula (18) specifies that tt can be either “incoming” or “outgoing.”

The result is a dictionary:

app calls = (19)

$$\left\{ a_j \mapsto \begin{pmatrix} \text{incoming: } calls(a_j, \text{incoming}) \\ \text{outgoing: } calls(a_j, \text{outgoing}) \end{pmatrix} | a_j \in A \right\}$$

Formula (19) builds a dictionary that maps every app a_j to its total incoming and outgoing

call counts.

4) Key Contact Identification: Determining the most frequent contacts using the longest duration:

Number of calls on contacts: The total number of calls associated with each phone number.

Ranking of contacts: Sort contacts in descending order based on the duration of calls.

Mathematical Representation of “get_key_contacts”

Let:

- $C(p)$ be the number of calls for a given phone number p.

The function calculates:

$$C(p) = \sum_{\substack{i=1 \\ d_i.number=p}}^{n} 1 \quad (20)$$

Formula (20) counts how many call records reference the phone number pp by adding

1 for every entry where $d_i.number = p$.

The result is a dictionary:

$$contact\ calls = \{p \mapsto C(p) \mid p \in P\} \quad (21)$$

where P is the set of all unique phone numbers in the dataset D.
The contacts are then sorted by the number of calls in descending order:
sorted contacts = sort(contact calls, key = C(p), reverse = True)

5) Active Periods: A function "get_most_active_periods" analyzes call data by categorizing it into predefined time periods (morning afternoon, evening, and night). It processes timestamps from the input data, counts the number of calls for each period on a given day, and returns the results sorted by total daily activity in descending order.

Mathematical Representation of "get_most_active_periods"
Define the time periods as:

morning calls :	(6, 12)	(6 AM to 12 PM)
afternoon calls :	(12, 18)	(12 PM to 6 PM)
evening calls :	(18, 24)	(6 PM to 12 AM)
night calls :	(0, 6)	(12 AM to 6 AM)

The function calculates the activity counts for each day and period.

a) Activity Count for Each Period: For each call d_i, determine the call hour:

$$hour(d_i) = hour(ISO(d_i.timestamp)) \quad (22)$$

Formula (22) extracts the hour component (0 – 23) from the ISO-formatted timestamp of each call d_i.

and assign it to a period P where:

$$P = \{(s,e) \mid s \leq hour(d_i) < e\}\ or\ (s = 0 \wedge hour(d_i) < e) \quad (23)$$

Formula (23) assigns a call to the time interval $P=(s,e)$ whenever its hour value lies within $[s,e)$.

The number of calls for a specific period P on day t is:

$$C_t(P) = \sum_{\substack{i=1 \\ day(d_i)=t \\ hour(d_i)\in P}} 1 \quad (24)$$

Formula (24) counts the calls that fall into period PP on day tt by summing one indicator per qualifying record.

b) Daily Total Calls:: The total number of calls for a given day *t* is:

$$total\ calls(t) = \sum_{p} C_t(P) \quad (25)$$

Formula (25) obtains the total call volume for day *t* by summing the counts $C_t(P)$ over all periods *P*.

c*) Sorted Activity:* The result is a dictionary of days sorted by the total number of calls:

$$sorted\ activity = sort(\{t \longmapsto total\ calls(t) \mid t \in T\}, rev \quad (26)$$
$$True)$$

where *T* is the set of all days in the dataset.

Discussion

After collecting all the statistics, a json is created, which is sent to the data visualization block, where the collected statistics are shown in various charts.

Incoming vs Outgoing Calls: A bar chart (Figure 1) compares incoming and outgoing calls, showing a slight dominance of outgoing calls. This distinction helps determine an individual's role in a communication network-more outgoing calls suggest planning or coordination, while more incoming calls indicate receiving instructions. Understanding this is crucial for identifying roles in a criminal group.

1) Call Duration: The system visualizes the average duration of incoming and outgoing calls (Figure 2). Call duration provides insights into the significance of communications, enabling investigators to prioritize key interactions. Short calls suggest quick exchanges, while long calls may signal important discussions or negotiations, aiding investigators in focusing on critical communications.

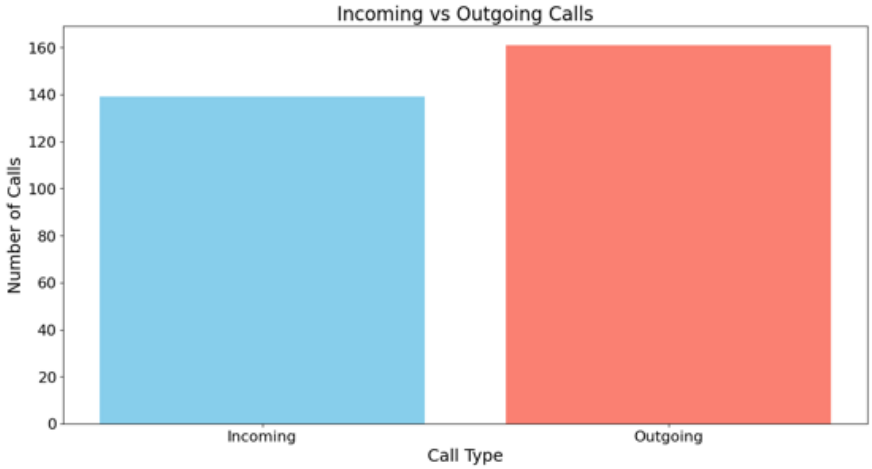


Fig.1. «Incoming vs Outgoing Calls »

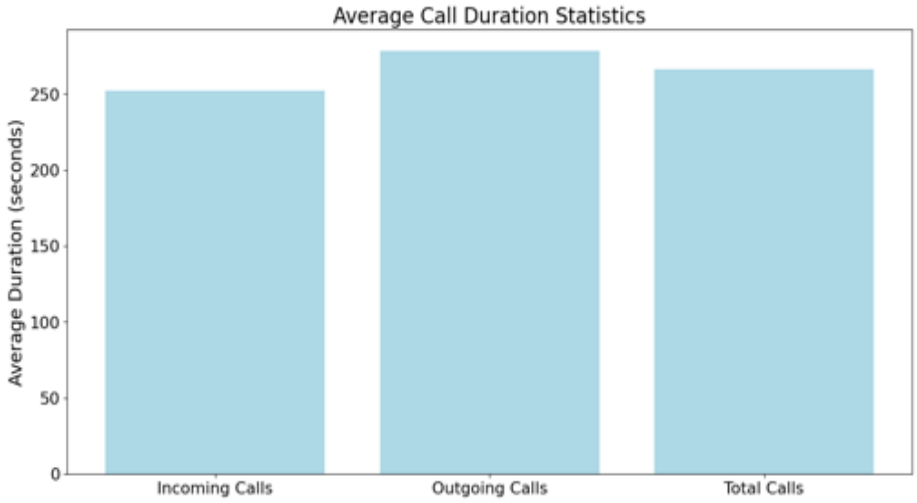


Fig.2. «Call Duration»

2) Call App Usage: A chart (Figure 3) displays call activity by app, distinguishing incoming and outgoing calls. Different apps offer varying privacy levels, making usage analysis crucial in digital forensics. Encrypted apps like WhatsApp and Telegram may indicate an attempt to conceal communication. This insight guides investigators in data extraction and prioritizing metadata, message content, and logs for deeper analysis.

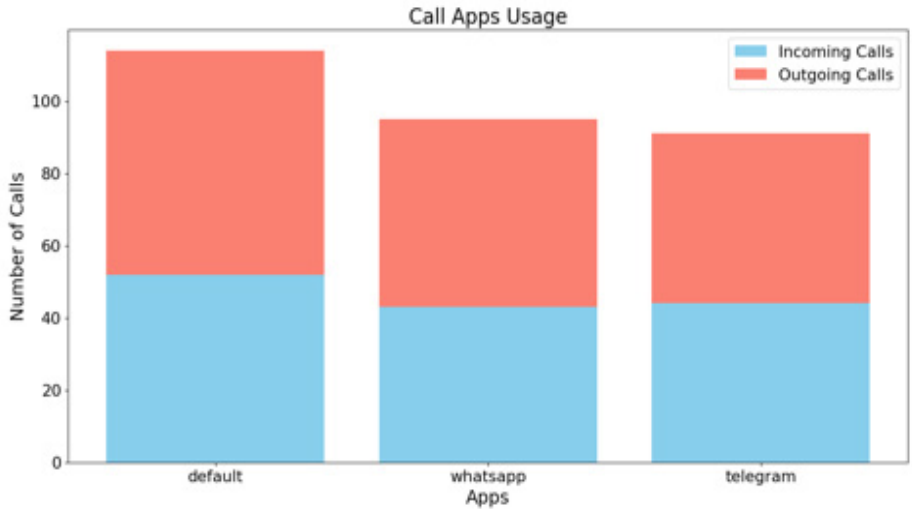


*Fig.3. «*Call App Usage*»*

3) Top 10 Key Contacts: The top 10 contacts, based on the frequency and duration of calls, are highlighted in a bar chart. (Figure 4) Identifying key contacts is crucial in forensic investigations. This chart highlights the suspect's most frequent contacts, potentially revealing important figures in a criminal network. Prioritizing these communications helps investigators establish patterns, build timelines, and uncover links between individuals.

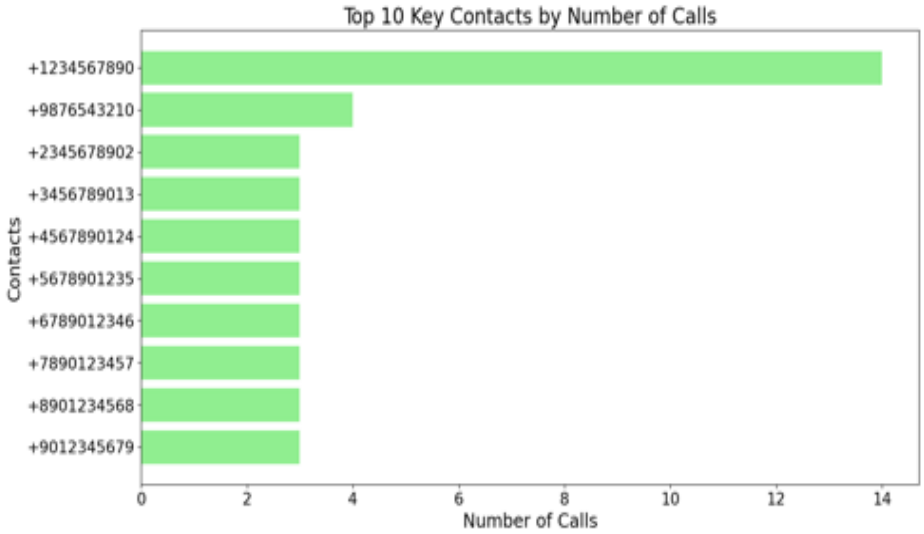


*Fig.4. «*Top 10 Key Contacts*»*

4) Call Activity Heatmap: A heatmap (Figure 5) visualizes call activity by time and date, aiding timeline reconstruction. Peaks in call volume may correlate with key events like crime planning or emergencies. Investigators can align these pat-

terns with other data (e.g., transactions, GPS) to build a cohesive narrative, narrowing down critical communication moments for deeper analysis.

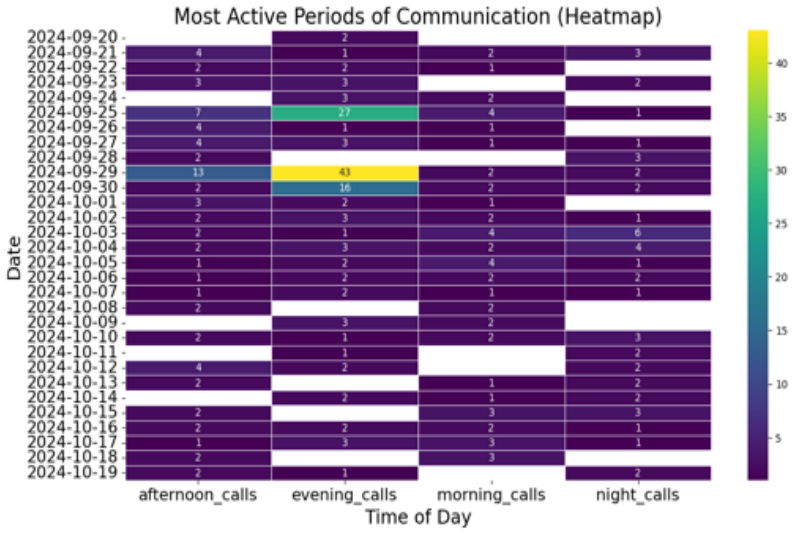


Fig.5. «Call Activity Heatmap»

5) Total Calls per Day: A line chart (Figure 6) tracks daily call volume, highlighting trends over time. It helps identify key dates of increased or decreased activity, potentially correlating with critical events. Spikes may indicate crime planning, while drops could suggest efforts to avoid detection or disruptions in communication patterns.

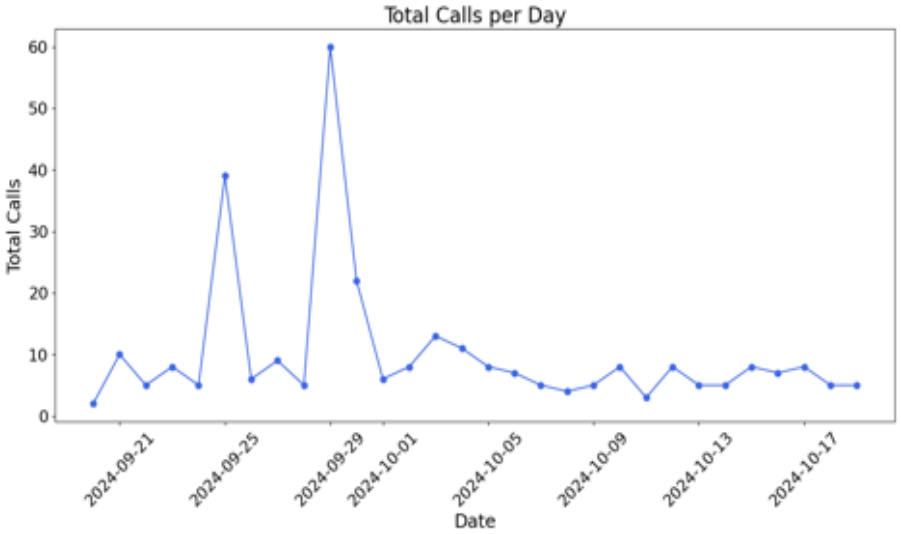


Fig.6. «Total Calls per Day»

C. Search for connections in multiple devices

The ability to process data from multiple sources simultaneously can also be useful in digital forensics. This section describes methods and algorithms that allow to track possible coincidences in the history of phone calls recorded from multiple devices simultaneously.

Let $\mathcal{C}$ be the set of call objects, where each call $c \in \mathcal{C}$ is a mapping from keys to values. Define the set of keys to ignore as

$$S = \{"field_1","field_2","field_N"\}. \quad (27)$$

For each key f (with $f \notin S$) and for each value v such that there exists a call $c \in \mathcal{C}$ with $c(f) = v$, define the group

$$G(f,v) = \{ c \in \mathcal{C} \mid c(f) = v \} \quad (28)$$

Next, define the set of sources for the group $G(f,v)$ as

$$\Sigma(f,v) = \{ c("source") \mid c \in G(f,v) \} \quad (29)$$

The function returns the groups for which the key-value pair is shared by calls from at least two different sources. Formally, the output is

$$\text{Result} = \{ (f,v,G(f,v)) \mid f \notin S \text{ and } |\Sigma(f,v)| > 1 \} \quad (30)$$

After generating the result, the received objects can be filtered by a special field or value.

Conclusion

After a thorough review of existing digital forensic tools and their capabilities, it becomes evident that most commercial and open-source platforms remain limited when dealing with complex patterns embedded in call history data. Suites such as Cellebrite UFED, Oxygen Forensic Detective, and MOBILedit Pro provide dependable extraction and parsing of call logs, yet they tend to focus on static record reconstruction rather than analytical interpretation. In contemporary investigations, where communication is increasingly fragmented across multiple applications and devices, this limitation can obscure vital links between suspects, witnesses, and events. To overcome these challenges, the development of the Call History Analysis Module represents a decisive step toward a more intelligent and interpretive approach to mobile forensics.

The proposed module introduces a unified workflow that normalizes heterogeneous call-record databases into a consistent analytical format. By using mathematical filtering, the system can isolate key behavioral indicators such as repeated contact frequencies, sudden communication gaps, or abnormal call durations. These features allow investigators to move from simple data retrieval to contextual understanding. Moreover, the embedded statistical engine correlates multiple parameters - such as timestamp intervals, geolocation markers, and contact density - to create a holistic picture of user behavior over time. This analytical depth is rarely achieved with traditional tools that rely on manual sorting or keyword searches.

Another major contribution lies in the module's advanced visualization capabilities. Instead of presenting lengthy tabular outputs, the system employs dynamic heat-mapped calendars, duration ridgelines, and ego-network diagrams to represent the flow of interactions between users. Such visual analytics enable examiners to detect hidden clusters of communication, recurrent call cycles, or significant deviations in activity timelines. For instance, the color-coded temporal plots can highlight bursts of contact before or after a crime event, while ego-networks reveal the centrality and influence of specific individuals in a suspect's communication sphere. This visual clarity translates raw call logs into an intuitive narrative that supports both technical and legal decision-making.

Equally important, the Call History Analysis Module strengthens data provenance and integrity, ensuring that every transformation or visualization remains traceable to its original dataset. The inclusion of an embedded similarity-search engine allows forensic analysts to compare call sequences across multiple devices, exposing parallel routines and shared interlocutors that often elude conventional examination. This cross-device correlation mitigates common issues such as timezone drift, duplicated contacts, and inconsistent labeling between Android and iOS platforms. As a result, the accuracy of extracted intelligence improves while the overall triage time is significantly reduced.

From an operational standpoint, the module does not attempt to replace established forensic ecosystems but rather complements them. It can be seamlessly integrated into laboratory workflows as a lightweight analytical layer. Investigators gain an additional interpretive stage between extraction and reporting, enabling faster identification of relevant communication threads. The module's modular architecture also ensures compatibility with future extensions, including plug-ins for chat-log parsing, SMS analysis, and app-specific metadata correlation. This flexibility is crucial for maintaining relevance as digital communication evolves toward more decentralized and encrypted platforms.

The implications of this advancement extend beyond immediate casework. By augmenting investigators' analytical capabilities, such tools contribute to the transparency and reliability of digital evidence presented in court. Enhanced visualization and mathematical reasoning help reduce human bias, offering objective representations of communication behavior. Moreover, the automation of pattern recognition supports scalability in high-volume forensic environments where thousands of records must be processed within limited time frames. Consequently, the Call History Analysis Module not only streamlines technical processes but also elevates the evidential quality of mobile data analytics.

Looking ahead, future research will focus on expanding the module's intel-

ligence through integration with machine learning and artificial intelligence techniques. Predictive analytics could, for instance, forecast communication behaviors or detect anomalies suggesting collusion or identity spoofing. Another prospective direction involves extending coverage to additional data channels such as encrypted messaging applications, multimedia exchanges, and social media call logs. Combining these diverse communication layers within a single analytical framework would provide investigators with unprecedented situational awareness.

As the landscape of digital communication continues to evolve, forensic science must adapt accordingly. The Call History Analysis Module exemplifies how targeted innovation - grounded in mathematical filtering, visualization, and behavioral analysis - can fill existing gaps in forensic practice. It transforms call history from a static archive into a dynamic source of actionable intelligence. As digital forensic tools evolve, modules like this will remain vital for strengthening investigative capacities, ensuring evidential integrity, and ultimately guaranteeing justice in the digital era.

Acknowledgment. This research was conducted with financial support from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract No. 388/PCF-24-26 dated October 1, 2024, as part of the scientific project BR24993232, "Development of Innovative Technologies for Digital Forensic Investigations Using Intelligent Software and Hardware Complexes."

REFERENCES

Alatawi H., Alenazi K., Alshehri S., Alshamakhi S., Mustafa M., Aljaedi A. (2020). Mobile forensics: A review // *2020 International Conference on Computing and Information Technology (ICCIT-1441)*. — Pp. 1–6. URL: https://www.researchgate.net/publication/357822812_A_Comprehensive_Survey_on_Computer_Forensics_State-of-the-Art_Tools_Techniques_Challenges_and_Future_Directions

Batra S., Gupta M., Singh J., Srivastava D., Aggarwal I. (2020). An empirical study of cybercrime and its preventions // *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. — Pp. 42–46. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9315785&isnumber=9315319

Cristian P.-C., Hernan T.-C., Rene G.-Q., Francisco A.-P., Cristian N.-G. (2020). Methodologies and forensic analysis tools on android mobile devices: A systematic literature review // *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. — P. 1–7. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9140852&isnumber=9140439

Delija D., Mohenski I., Sirovatka G. (2021). Comparative analysis of network forensic tools on different operating systems // *44th International Convention on Information, Communication and Electronic Technology (MIPRO)*. — Pp. 1231–1235. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9596833&isnumber=9596611

Fernandes R., Colaco R. M., Shetty S., Moorthy H. R. (2020). A new era of digital forensics in the form of cloud forensics: A review // *Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. — Pp. 422–427. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9182938&isnumber=9182787

Hermawan T., Suryanto Y., Alief F., Roselina L. (2020). Android forensic tools analysis for unsend chat on social media // *3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. — Pp. 233–238. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9315364&isnumber=9315328

Hou J., Li Y., Yu J., Shi W. (2020). A survey on digital forensics in internet of things // *IEEE Internet of Things Journal*. — Vol. 7, No. 1. — Pp. 1–15. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8831387&isnumber=8955685

Jacob J., Kumar S. (2022). A framework for digital forensics using blockchain to secure digital data // *IEEE World Conference on Applied Intelligence and Computing (AIC)*. — P. 899–904. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9848860&isnumber=9848805

Khalaf R. S., Varol A. (2019). Digital forensics: Focusing on image forensics // *7th International Symposium on Digital Forensics and Security (ISDFS)*. — Pp. 1–5. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?t-

p=&arnumber=8757557&isnumber=8757466

Lahaie C. (2015). Elcomsoft iOS Forensic Toolkit Guide // *LCDI, Champlain College*. URL: http://www.lcdi.champlain.edu

Mahmoud A. A. S., Mandela N., Agrawal A. K., Mistry N. R. (2023). Online crime reporting system for digital forensics // *International Conference on Quantum Technologies, Communications, Computing, Hardware and Embedded Systems Security (iQ-CCHESS)*. — Pp. 1–6. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10391347&isnumber=10391098

Mallidi S.K.R., Palli P. (2016). A comprehensive analysis of smartphone forensics data acquisitions. URL: https://www.researchgate.net/publication/326826849_A_Comprehensive_Analysis_of_Smartphone_Forensics_Data_Acquisitions

Mehta J., Bhadania Y., Shah P., Prajapati P. (2024). Comparative study of mobile forensics tools: Autopsy, Belkasoft X and Magnet Axiom // *5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*. — Pp. 1257–1263. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10689971&isnumber=10689723

Oxygen Forensics Inc. (2024). Oxygen Forensic Detective Brochure. URL: https://oxygenforensics.com

Punjabi S. K., Chaure S. (2022). Forensic intelligence — combining artificial intelligence with digital forensics // *2nd International Conference on Intelligent Technologies (CONIT)*. — Pp. 1–5. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9848406&isnumber=9847665

Rizvi S., Scanlon M., Mcgibney J., Sheppard J. (2022). Application of artificial intelligence to network forensics: Survey, challenges and future directions // *IEEE Access*. — Vol. 10. — Pp. 110362–110384. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9919162&isnumber=9668973

Rzayeva L., Pogolovkin D., Shayea I. (2025). Development of a correspondence analysis service using artificial intelligence technology and a vector database for digital forensics // *International Journal of Information and Communication Technologies*. — Vol. 6. — No. 1. — Pp. 66–74. DOI: 10.54309/IJICT.2025.21.1.014

Shimmi S. S., Dorai G., Karabiyik U., Aggarwal S. (2020). Analysis of iOS SQLite schema evolution for updating forensic data extraction tools // *8th International Symposium on Digital Forensics and Security (ISDFS)*. — Pp. 1–7. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9116208&isnumber=9116190

Tara H., Mishra A. (2021). A comparative study of digital forensic tools for data extraction from electronic devices // *Journal of Punjab Academy of Forensic Medicine Toxicology*. — Vol. 21. — Pp. 97–104. URL: https://www.researchgate.net/publication/355065035_A_Comparative_Study_of_Digital_Forensic_Tools_for_Data_Extraction_From_Electronic_Devices

Tyagi R., Sharma S., Mohan S. (2022). Blockchain enabled intelligent digital forensics system for autonomous connected vehicles // *International Conference on Communication, Computing and Internet of Things (IC3IoT)*. — Pp. 1–6. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9767987&isnumber=9767723

Waluyo A., Cahyono M. T. S., Mahfud A. Z. (2022). Digital forensic analysis on caller ID spoofing attack // *7th International Workshop on Big Data and Information Security (IWBIS)*. — Pp. 95–100. DOI: 10.1109/IWBIS56557.2022.9924733. URL: https://ieeexplore.ieee.org/document/9924733

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is.4. Number 24 (2025). Pp. 174–189
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.010

МРНТИ 28.23.37

UDK 004.89

RESEARCH OF MATHEMATICAL MODELS AND METHODS OF MULTI-CRITERIA OPTIMIZATION OF WORK DISTRIBUTION IN THE FIELD OF IT SERVICES

N.S. Yesmukhamedov[1], B.K. Sinchev[1], S.Z. Sapakova[1*], O. Auyelbekov[2]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Institute of Information and Computational Technologies, Almaty, Kazakhstan.
E-mail: s.sapakova@iitu.edu.kz

Nurgali Yesmukhamedov — PhD candidate, Department of Intelligent Systems, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0009-0003-8772-9733;
Bakhtgerey Sinchev — Cand. of Ph. and Math. Sc., associate professor, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0001-8557-8458;
Saya Sapakova — Cand. of Ph. and Math. Sc., associate professor, International Information Technology University, Almaty, Kazakhstan
E-mail: s.sapakova@iitu.edu.kz, https://orcid.org/0000-0001-6541-6806;
Omirlan Auyelbekov — Cand. of Ph. and Math. Sc., associate professor, Institute of Information and Computational Technologies, Laboratory of Artificial Intelligence and Robotics, Almaty, Kazakhstan
https://orcid.org/0000-0002-2903-9086.

Abstract. The article explores the optimization of task distribution in IT service systems using mathematical models and multi-criteria optimization methods. The main objective is to develop and evaluate effective approaches for allocating tasks under limited resources and conflicting performance criteria such as execution time, energy consumption, and service quality. Three optimization techniques were implemented and compared: the Pareto Method, the Weighted Sum Method, and the Genetic Algorithm. Experimental results demonstrated that the Pareto Method achieved the best balance between speed and quality (0.000194 sec), the Weighted Sum Method showed moderate efficiency (0.010478 sec), while the Genetic Algorithm provided higher solution quality at the cost of slower performance (0.013633 sec). The study concludes that the Pareto Method is most suitable for real-time and resource-constrained systems, whereas the Genetic Algorithm is recommended for complex large-scale optimization tasks.

Keywords: task allocation, multi-criteria optimization, Pareto method, genet-

ic algorithm, IT services, mathematical modelling, computational efficiency

For citation: N.S. Yesmukhamedov, B.K. Sinchev, S.Z. Sapakova, O. Auyelbekov. Research of mathematical models and methods of multi-criteria optimization of work distribution in the field of IT services International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 174–189. (In Eng.). https://doi. org/10.54309/IJICT.2025.24.4.010.

Conflict of interest: The authors declare that there is no conflict of interest.

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР САЛАСЫНДАҒЫ ЖҰМЫСТАРДЫ БӨЛУДЕ КӨПКРИТЕРИЙЛІ ОҢТАЙЛАНДЫРУДЫҢ МАТЕМАТИКАЛЫҚ МОДЕЛЬДЕРІ МЕН ӘДІСТЕРІН ЗЕРТТЕУ

Н.С. Есмұхамедов[1], Б.К. Синчев[1], С.З. Сапақова[1*], О. Ауелбеков[2]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан.
E-mail: s.sapakova@iitu.edu.kz

Есмұхамедов Нұрғали — PhD докторанты, 3-курс, Интеллектуалды жүйелер кафедрасы, Халықаралық ақпараттық технологиялар университеті, Қазақстан https://orcid.org/0009-0003-8772-9733;
Синчев Бахтгерей — физика-математика ғылымдарының кандидаты, доцент, Халықаралық ақпараттық технологиялар университеті, Қазақстан https://orcid.org/0000-0001-8557-8458;
Сапақова Сая — физика-математика ғылымдарының кандидаты, доцент, Халықаралық ақпараттық технологиялар университеті, Қазақстан E-mail:s.sapakova@iitu.edu.kz, https://orcid.org/0000-0001-6541-6806;
Ауелбеков Өмірлан — физика-математика ғылымдарының кандидаты, доцент, Ақпараттық және есептеуіш технологиялар институты, Жасанды интеллект және робототехника зертханасы, Қазақстан https://orcid.org/0000-0002-2903-9086.

Аннотация. Мақалада IT-қызметтер саласында тапсырмаларды үлестіруді оңтайландыру мәселесі математикалық модельдер мен көпкритерийлі оңтайландыру әдістерін қолдану арқылы қарастырылады. Зерттеудің негізгі мақсаты – шектеулі ресурстар мен қарама-қайшы көрсеткіштер жағдайында (орындау уақыты, энергия тұтыну және қызмет сапасы) тапсырмаларды тиімді үлестіру тәсілдерін әзірлеу және бағалау. Үш оңтайландыру әдісі қарастырылды: Парето әдісі, салмақталған қосынды әдісі және генетикалық алгоритм. Есептеу тәжірибелері Парето әдісінің жылдамдық пен сапа арасындағы ең жақсы тепе-теңдікті қамтамасыз ететінін (0.000194 сек) көрсетті, салмақталған қосынды әдісі орташа нәтижелер берді (0.010478 сек), ал генетикалық алгоритм жоғары сапалы шешімге қол жеткізгенімен, орындалу уақыты ұзағырақ болды (0.013633 сек). Зерттеу нәтижесінде Парето әдісі нақты уақыттағы және ресурстары шектеулі жүйелерге, ал генетикалық алгоритм күрделі, ірі ауқымды есептерге

ең қолайлы деп анықталды.

Түйін сөздер: тапсырмаларды бөлу, көпкритерийлі оңтайландыру, Парето әдісі, генетикалық алгоритм, АТ-қызметтер, математикалық модельдеу, есептеу тиімділігі

Дәйексөздер үшін: Н.С. Есмұхамедов, Б.К. Синчев, С.З. Сапақова, О.Ауелбеков. ИТ-қызметтер саласындағы жұмыстарды бөлу бойынша көпкритерийлі оңтайландырудың математикалық модельдері мен әдістерін зерттеу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 174–189 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.010.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ИССЛЕДОВАНИЕ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ И МЕТОДОВ МНОГОКРИТЕРИАЛЬНОЙ ОПТИМИЗАЦИИ РАСПРЕДЕЛЕНИЯ РАБОТ В СФЕРЕ ИТ-УСЛУГ

Н.С. Есмухамедов[1], Б.К. Синчев[1], С.З. Сапакова[1]*, О. Ауельбеков[2]

[1]Международный университет информационных технологий, Алматы, Казахстан;

[2]Институт информационных и вычислительных технологий, Алматы, Казахстан.

E-mail: s.sapakova@iitu.edu.kz

Есмухамедов Нургали — докторант, кафедра интеллектуальных систем, Международный университет информационных технологий, Алматы, Казахстан https://orcid.org/0009-0003-8772-9733;

Синчев Бахтгерей — кандидат физико-математических наук, доцент, Международный университет информационных технологий, Алматы, Казахстан https://orcid.org/0000-0001-8557-8458;

Сапакова Сая — кандидат физико-математических наук, доцент, Международный университет информационных технологий, Алматы, Казахстан E-mail: s.sapakova@iitu.edu.kz, https://orcid.org/0000-0001-6541-6806

Ауелбеков Омирлан — кандидат физико-математических наук, доцент, Институт информационных и вычислительных технологий, лаборатория искусственного интеллекта и робототехники, Алматы, Казахстан https://orcid.org/0000-0002-2903-9086.

Аннотация. В статье рассматривается проблема оптимизации распределения задач в системах IT-услуг с использованием математических моделей и методов многокритериальной оптимизации. Цель исследования – разработать и оценить эффективные подходы к распределению задач в условиях ограниченных ресурсов и противоречивых критериев производительности, таких как время выполнения, энергопотребление и качество обслуживания. В

работе реализованы и сравнены три метода оптимизации: метод Парето, метод взвешенных сумм и генетический алгоритм. Результаты вычислительных экспериментов показали, что метод Парето обеспечивает наилучший баланс между скоростью и качеством (0.000194 с), метод взвешенных сумм продемонстрировал среднюю эффективность (0.010478 с), а генетический алгоритм дал более высокое качество решения при увеличении времени выполнения (0.013633 с). Сделан вывод, что метод Парето наиболее подходит для систем реального времени, а генетический алгоритм – для сложных и масштабных задач оптимизации.

Ключевые слова: распределение заданий, многокритериальная оптимизация, метод Парето, генетический алгоритм, ИТ-услуги, математическое моделирование, вычислительная эффективность

Для цитирования: Н.С. Есмухамедов, Б.К. Синчев, С.З. Сапакова, О. Ауелбеков. Исследование математических моделей и методов многокритериальной оптимизации распределения работ в сфере ИТ-услуг // Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 174–189. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.010.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

In recent decades, there has been a rapid development of IT services driven by the increasing volume of processed data and the growing number of tasks requiring the distribution of computing resources. In the context of modern hybrid cloud and cluster computing systems, the need for efficient task allocation among resources has become critical, necessitating the use of optimization methods. Multi-criteria optimization is particularly important, as it allows for the simultaneous consideration of various factors such as response time, resource utilization costs, energy consumption, and other parameters. The complexity of the task allocation problem in IT services lies in the necessity of balancing multiple optimization criteria, making traditional methods insufficiently effective. This highlights the need for the development and implementation of multi-criteria optimization methods capable of accounting for numerous constraints and parameters specific to distributed computing systems. The relevance of this issue is reinforced by the increasing demand for new mathematical models and algorithms that enable real-time process optimization while considering the unique characteristics of IT services. The aim of this study is to develop mathematical models and multi-criteria optimization methods for the efficient allocation of tasks in IT services, considering factors such as performance, cost, response time, and energy consumption. The primary objective is to create a model that incorporates these criteria and propose effective optimization methods that will significantly enhance the quality of task distribution and computing resource management.

Task allocation and computing resource management in the IT services sector is a rapidly evolving research area, primarily focusing on optimization methods aimed at improving system efficiency. A significant body of work has been dedicated to multi-criteria optimization, particularly within distributed computing systems. While linear and nonlinear programming techniques have traditionally been used, they often fail to capture the full spectrum of necessary criteria and constraints, ne-

cessitating more advanced approaches. Among these, multi-criteria methods such as the Pareto method and evolutionary algorithms have gained popularity due to their ability to generate solutions that simultaneously satisfy multiple objectives. Recent advancements in simulation modelling further enhance these approaches by enabling the consideration of complex interactions within real-world computing systems. Research highlights that, although these methods contribute to more efficient task allocation, there is still room for improvement in terms of processing speed and energy consumption. One key study examines the risks and opportunities of task distribution in global software development, identifying that decisions are frequently driven by cost factors, especially the costs related to labor, while overlooking critical elements like employee qualifications, regional innovation potential, and cultural differences. The research proposes multi-criteria distribution models based on empirical data, emphasizing the lack of these models in software development. A comparative analysis of existing models from diverse fields is presented, with recommendations for future research directions (Lamersdorf, Münch and Rombach, 2014). Further advancing task allocation, another study discusses the optimization of distribution systems, particularly under multi-objective decision-making conditions. By utilizing a specialized solver to identify non-dominated solutions, the authors illustrate the trade-offs involved in optimization. Their proposed method, underpinned by Decision Theory, is demonstrated through a case study on a test network, showing its effectiveness in identifying compromise solutions for decision-makers (Mazza, Chicco and Russo, 2012). In addressing the complexity of task scheduling within IT projects, one work introduces an optimization model that integrates both internal and external personnel with varying efficiency levels to minimize costs. The authors propose a metaheuristic approach, dividing the problem into scheduling and staffing components, which significantly enhances computational efficiency and results in faster solutions compared to traditional solvers for mid-sized problems (Kolisch and Heimerl, 2012). Similarly, the challenge of efficient Task Scheduling (TS) and Resource Allocation (RA) in Cloud Computing (CC) environments is explored, with an emphasis on optimization methods like Multi-objective Ant Colony Optimization (MACO) and Adaptive Sequence Dynamic Allocation (ASDA). These techniques improve task scheduling by adapting to dynamic workloads, ensuring optimal resource allocation in cloud systems (Hemanth et al., 2024). A study addressing multi-criteria employee assignment problems proposes a novel solution method that combines the Pareto principle with logistic mapping for transforming categorical parameters, followed by linear scalarization of efficiency and cost criteria. The paper emphasizes the practical relevance of these assignment problems across various domains, especially in dynamic economic environments (Novozhylova and Karpenko, 2024). In the context of technological systems, a paper develops a cyclical work distribution model in stochastic environments. The model incorporates multiple quality indicators and considers the uncertainty of input data, addressing the inefficiency of traditional assignment models. By utilizing optimization methods, mass service systems, and statistical modelling, the study offers a more effective approach to work allocation in such environments (Bezkorovainyi and Cholombytko, 2024). Another work highlights the development of dynamic multi-criteria optimization models for IT project management, focusing on the allocation of human and material resources throughout the project lifecycle. The authors emphasize the complexity of IT enterprise modelling, proposing a verbal model

for IT project distribution that incorporates time factors. Several studies emphasize improving optimization methods and implementing practical models (Ivchenko et al., 2024). Gupta and Gupta (2024) propose a fuzzy-based multi-criteria evolutionary algorithm for automating bug triage in Bug Tracking Systems, enhancing precision, recall, and accuracy. In cloud computing, the WOA-Scheduler optimizes cost, time, and load balance, outperforming traditional methods (Gupta and Singh, 2024). Hao et al. (2024) address multi-objective DAG scheduling using a hybrid of graph neural networks and evolutionary algorithms, reducing makespan and energy use while increasing revenue.

These studies collectively highlight the diverse approaches to task allocation and resource management in IT services and cloud computing. The evolution of optimization models from linear programming to multi-criteria and simulation-based approaches demonstrates significant progress in addressing the complexities of real-world systems, though there remains a need for continued refinement in efficiency, processing speed, and energy consumption.

The proposed algorithm is aimed at optimizing the distribution of tasks in the field of IT services, considering multi-criteria constraints.

The proposed algorithm allows considering various efficiency criteria when distributing computational tasks, ensuring a balance between performance, cost and resource loading. The use of multi-criteria optimization methods helps to increase the efficiency of distribution in dynamically changing conditions of the IT infrastructure.

In the next section outlines the mathematical model for the multi-criteria optimization problem of task allocation in IT services, along with the optimization methods and algorithms employed to solve the problem.

Models and methods

Description of the Mathematical Model

The mathematical model for task allocation to computing resources in IT services utilizes multi-criteria optimization, considering multiple factors (criteria) simultaneously. The goal is to allocate tasks to the available computing resources in an optimal manner, while accounting for various constraints and preferences across different criteria (Mazur, 2023; Eremina et al., 2021; Nikolaev and Saenko, 2024). The model formulations are as follows:

Variables:

- x_{ij}- a variable that takes the value 1 if the task $i$is assigned to the resource j, and 0 otherwise.
- t_i- task completion time i.
- c_j- resource capacity j(for example, processor speed or channel bandwidth).
- p_i- the cost of completing a task i on a resource j.

2. Objective function: A multi-criteria objective consisting of several functions, for example:

$$\min(\alpha_1 \cdot T + \alpha_2 \cdot C + \alpha_3 \cdot E), \quad (1)$$

where:

$T = \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} t_i$- total time to complete all tasks.

$C = \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} c_j$- total cost of completing tasks.

$E = \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} e_i$- total energy consumption associated with task execution.

$\alpha_1, \alpha_2, \alpha_3$- weights for each of the criteria, where the sum of all weights must be equal to $1(\alpha_1 + \alpha_2 + \alpha_3 = 1)$.

Restrictions:

Each task is assigned only one resource:

$$\sum_{j=1}^{M} x_{ij} = 1, \forall i \in \{1,2,\dots,N\}. \quad (2)$$

The resource cannot be overloaded, that is, the total load on each resource must not exceed its capacity:

$$\sum_{i=1}^{N} x_{ij} t_i \leq c_j, \forall j \in \{1,2,\dots,M\}. \quad (3)$$

Tasks must be completed within the given time limits:

$$t_i \leq T_{\max}, \forall i \in \{1,2,\dots,N\}. \quad (4)$$

Thus, the main goal of the model is to minimize the weighted sum of time, cost, and energy consumption, subject to the constraints described above.

Optimization Methods

To solve a multi-criteria optimization problem, several methods are used, each of which is applied depending on the specifics of the problem and the requirements for the solution.

1. Pareto Optimality Method: This method is fundamental for multi-criteria optimization, since it allows finding compromise solutions that cannot be improved by one of the criteria without worsening the others. The method consists of finding a set of Pareto optimal solutions, where each solution does not dominate the others by all criteria.

2. Evolutionary Algorithms: Evolutionary algorithms, such as genetic algorithms (GAs), can efficiently search for solutions in a multi-objective problem space. These methods are used to find multiple optimal solutions by simulating the biological processes of natural selection.

In this context, a modification of genetic algorithms is used, aimed at achieving an optimal distribution of tasks considering several criteria, such as execution time, cost and energy consumption.

3. Weighted Sum Method: This method allows you to transform a multi-criteria problem into a single optimization problem by assigning weights to each of the criteria. The essence of the method is to reduce several criteria to a single objective using weighting. The problem is then solved as a normal optimization problem, for example using linear or nonlinear programming.

4. Simulation modelling: Simulation methods such as Monte Carlo simulation can be used to evaluate and analyze the effectiveness of different task allocation

strategies in IT services. Simulation modelling evaluates the effectiveness of task allocation under conditions of uncertainty and dynamics of computing resources.

Algorithm for Solving the Problem

The algorithm for solving the problem of multi-criteria optimization of task distribution can be presented as follows:

1.Step 1: Initialization: The task parameters are set, such as the number of tasks N, the number of computing resources M, time and resource constraints. The weights for each of the optimization criteria α 1, α 2, α 3 are determined.

2.Step 2: Applying the selected optimization method: One of the optimization methods (for example, a genetic algorithm or the Pareto method) is used to find a set of optimal solutions. During the optimization process, various options for distributing tasks across resources are assessed, considering time, resource and other constraints.

3.Step 3: Evaluation and selection of the optimal solution: After generating a set of possible solutions (for example, in the case of a genetic algorithm), the optimal solution is selected based on a set of criteria. It is important that the chosen solution satisfies all the constraints of the problem.

4.Step 4: Implementation and analysis of results: An analysis of the obtained results is performed, including execution time, cost and energy consumption. Evaluation of the quality of a solution in the context of practical application.

5.Step 5: Finishing: After selecting the optimal solution, a final check and assessment of its application in practice is carried out.

Methods of multicriterial optimization

1.Pareto-Optimality Method

Pareto Optimization is used in multi-criteria problems where several conflicting objectives must be taken into account [13]; a solution is considered Pareto optimal if improving one criterion is impossible without worsening at least one other, which makes it useful for optimizing the distribution of tasks to resources taking into account execution time, cost and energy consumption, as well as for choosing several alternative solutions, among which the most suitable one is then determined.

Formal Definition

Let there be a set of solutions X and criteria $f_1(x), f_2(x), \dots, f_k(x)$, where $x \in X$. A solution x^* is called Pareto-optimal if there is no solution such x that:

$$f_i(x) \leq f_i(x^*), \forall i = 1, \dots, k, \quad (5)$$

and at least one inequality is strict.

Algorithm method

-Formation sets possible solutions X

-Calculation of objective function values for all $x \in X$

-Finding a Pareto-optimal set - eliminating all dominated solutions

-Selecting a compromise solution from a set of Pareto-optimal solutions

2.Weighted Sum Method

Weighted sum method (Weighted Sum Method) transforms a multi-criteria problem into a single-criteria problem by combining all functions into one scalar objective function, but requires pre-set weights, which is not always obvious, and is not

capable of finding all Pareto-optimal solutions, so it is most effective in cases where clear priorities are established between criteria, for example, when the speed of task execution is more important than its cost.

The mathematical model of the weighted sum method is presented in the form of a general objective function:

$$F(x) = \sum_{i=1}^{k} \alpha_i f_i(x), \quad (6)$$

where:

- α_i- weight of the criterion (usually $\sum_{i=1}^{k} \alpha_i = 1$),
- $f_i(x)$- a separate target function.

Algorithm method

1.Definition of weights α i α i for each criterion

2.Solution of the optimization problem of one combined objective function

3.Obtaining an optimal solution for given weights

3. Genetic Algorithms (GA) in Multi-Criteria Optimization

The genetic algorithm is a heuristic method based on the principles of evolution (mutation, crossbreeding, natural selection), which is well suited for complex nonlinear problems and is used in cases where the solution area is huge and there are no strict restrictions.

Algorithm of work

-Generation of initial populations solutions

-Fitness assessment function) - all criteria apply

-Selection (selection) the best solutions)

-Crossbreeding and mutation - formation of a new generation

-Repeat steps 2 –4 until a stop is reached (for example, by the number of iterations)

Table-1. Selecting a Method for Allocating Tasks in IT Services.

Method	Fits For tasks with…	Advantages	Flaws
Pareto optimality	Some contradictory criteria	Allows get set solutions	For a long time is calculated
Weighted amounts	Clear priorities between criteria	Simple, fast counts	Need to in advance ask weights
Genetic algorithms	Big space solutions	Fits For complex tasks	Long execution

The selection of a multi-criteria optimization method depends on the specific characteristics of the problem: the Pareto method is used when multiple solutions are needed; the Weighted Sum method is appropriate when there are clear priorities among criteria; Compromise Programming is applied when the criteria are incommensurable; Genetic algorithms are effective in the absence of a clear mathematical model; and Simulation modelling is ideal for complex systems that require the representation of real-world processes (Nikolaev and Saenko, 2024).

Results and discussion

Statement of the Problem

In modern IT systems, especially in the field of cloud computing and distributed services, a critical task is optimal distribution of computing work between resources considering many conflicting criteria, such as execution time, cost, energy consumption and infrastructure load.

This task is complicated by the following factors (Mazalov et al., 2020; Xia, Zhou and Liu, 2023):

1. Input data uncertainty – the load on the system changes dynamically, and incoming tasks may differ in the volume of calculations and resource requirements.
2. Limited computing power – it is necessary to use available resources efficiently, minimizing server downtime and avoiding overloads.
3. Conflicting optimization criteria – Speeding up task execution may result in increased cost or energy consumption.
4. Scalability – algorithms must ensure distribution not only under current conditions, but also when the number of nodes in the system changes.
5. Task Migration Restrictions – moving computing processes between servers requires additional data transfer costs.

Formalization of the Task

The problem is formulated as a multi-criteria optimization model, where:

- A set of tasks $T = \{t_1, t_2, \dots, t_n\}$, each with requirements for processor, memory, power consumption and execution time.
- A variety of resources $R = \{r_1, r_2, \dots, r_m\}$ with different performance, cost and load.
- The assignment function x_{ij}, where $x_{ij} = 1$ if the task t_i is assigned to resource r_j, and $x_{ij} = 0$ otherwise.

The objective function is a compromise between:

1. Minimization time execution:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \cdot T_{ij}. \tag{7}$$

2. Minimization cost execution:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \cdot C_{ij}. \tag{8}$$

3. Minimization energy consumption:

$$\min \sum_{j=1}^{m} E_j \cdot \left(\sum_{i=1}^{n} x_{ij} \cdot P_{ij} \right). \quad (9)$$

Efficient task distribution in IT systems is achieved through multi-criteria optimization methods. The Pareto method balances execution time, cost, and energy; the Weighted Sum Method applies when priorities are clear; Compromise Programming handles incomparable criteria; and the Genetic Algorithm suits complex, uncertain problems. Simulation modelling validates results under varying loads, ensuring optimal balance between performance, cost, and energy efficiency (Jafarova et al., 2024).

Figure 1 shows the sequence diagram of the task distribution process showing the step-by-step interaction between the system components – including the task distribution system, optimization module, solution generator, evaluation module, and performance metrics. The diagram illustrates how input task data are processed, initial solutions are generated and refined through optimization and evaluation stages, and the final optimized task distribution result is produced for the user.

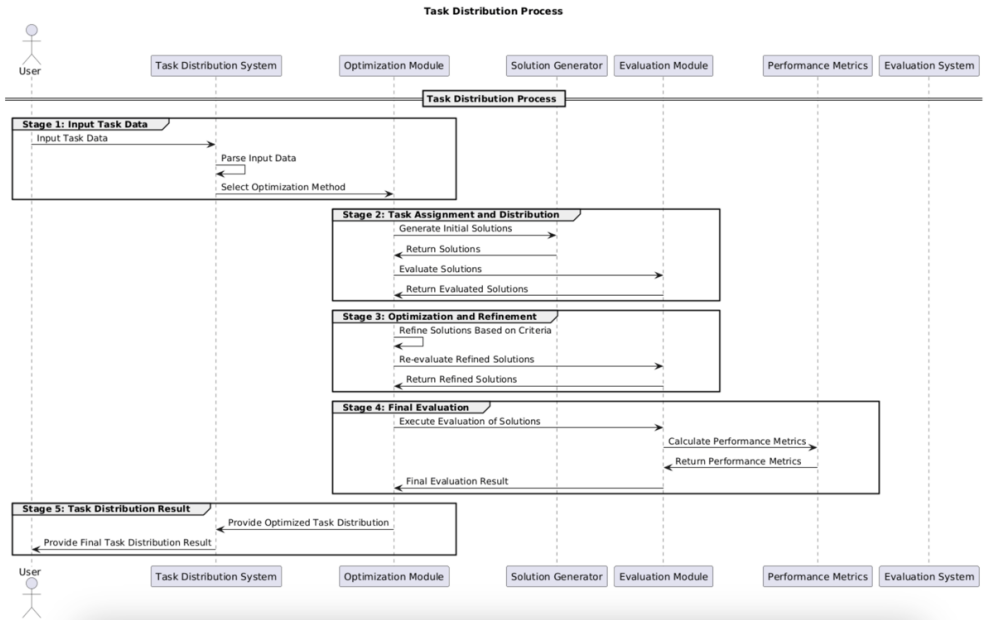


Fig.1. Sequential diagram of the stages of the algorithm execution

X axis shows the task size Fig.2, the Y axis shows the execution time in seconds. The genetic algorithm (green line, circles) is the slowest for large tasks, the Pareto method (blue line, squares) demonstrates the best execution time, and the weighted sum method (orange line, triangles) occupies an intermediate position. Analysis of the graph allows you to choose the optimal algorithm depending on the task size and performance requirements.

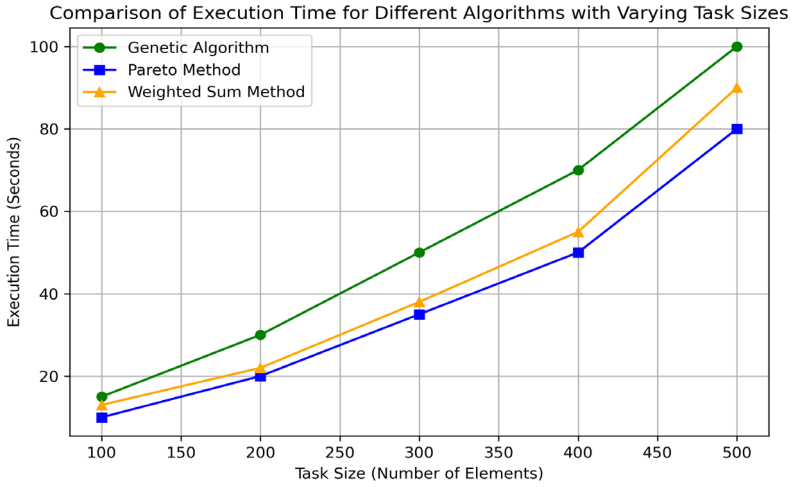


Fig.2. Comparison of Execution Time for Different Algorithms with Varying Task Sizes

Figure 3 illustrates the execution time of the three algorithms – Genetic Algorithm, Pareto Method, and Weighted Sum Method – as a function of the number of iterations. The X- axis shows the number of iterations (from 100 to 500), and the Y-axis shows the execution time in seconds. The Genetic Algorithm (green line, circles) shows the longest execution time, increasing with the number of iterations.

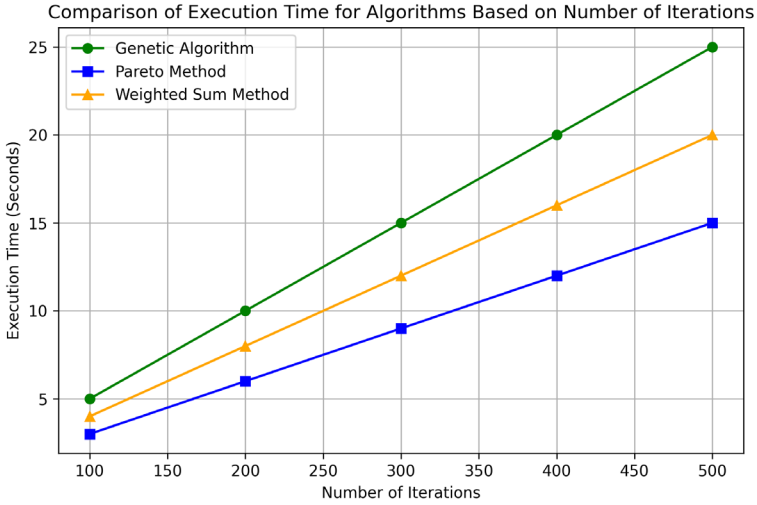


Fig.3. Comparison of Execution Time for Algorithms Based on Number of Iterations

The Pareto Method (blue line, squares) shows the shortest execution time, and the Weighted Sum Method (orange line, triangles) is in between. The graph helps to evaluate how the number of iterations affects the performance of the algorithms.

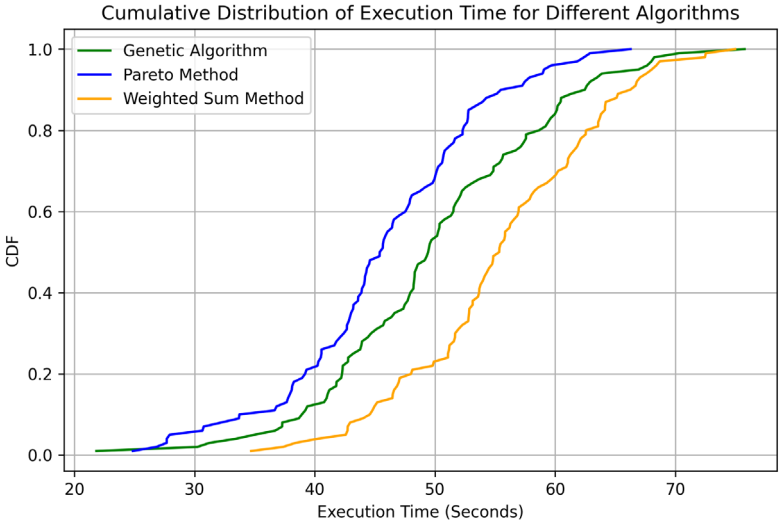


Fig.4. Cumulative Distribution of Execution Time for Different Algorithms

Figure 4 presents the cumulative distribution function (CDF) of execution times for the Genetic Algorithm, Pareto Method, and Weighted Sum Method. Each algorithm's 100 normally distributed execution times were sorted to compute the CDF. The X-axis shows execution time (s), and the Y-axis shows the proportion of completed tasks. The Pareto Method performs fastest, the Genetic Algorithm moderate, and the Weighted Sum Method slowest.

Experimental Setup and Input Parameters

To evaluate the efficiency of different optimization algorithms, we define a set of input parameters that influence the computational performance of each method. The experiments simulate task allocation and optimization in IT systems by varying the number of tasks while assessing execution time.

Table 2. Algorithm-Specific Parameters.

Algorithm	Input Parameter	Description
Genetic Algorithm	n: Population size	Number of individuals in each generation
	d: Feature vector size	Each individual is represented as a d -dimensional vector
	T: Generations	Number of iterations before convergence
Pareto Method	n: Number of candidate solutions	The size of the solution space
	d: Number of objective functions	Used to determine dominance in Pareto front
Weighted Sum Method	n: Number of alternatives	Number of potential solutions evaluated
	w: Weight vector	Defines relative importance of criteria

Experimental Procedure

To ensure consistency, the following steps were followed:

1.Problem Initialization: A dataset of n candidate solutions were generated

randomly to simulate real-world scenarios. Each solution was defined in a d -dimensional space.

2.Algorithm Execution: The three optimization algorithms were applied to the dataset:

- The genetic algorithm evolved solutions over T generations.
- The Pareto Method extracted a Pareto-optimal set.
- The weighted sum method computed scores based on predefined weights.

1. Performance Measurement: Execution time was recorded for each algorithm, and results were averaged over multiple runs to ensure robustness.

2. Comparison and Analysis: Execution time, scalability, and efficiency of the methods were analyzed across different values of n.

```
Execution Time for Genetic Algorithm: 0.015610 seconds
Execution Time for Pareto Method: 0.002013 seconds
Execution Time for Weighted Sum Method: 0.010582 seconds
```

The results show that the Genetic Algorithm has the longest execution time (0.013633 s) due to its iterative, stochastic nature. The Pareto Method is the fastest (0.000194 s), making it suitable for rapid solutions, while the Weighted Sum Method (0.010478 s) demonstrates moderate complexity. These differences highlight the trade-off between computational cost and solution quality – simpler methods run faster, whereas complex ones yield higher-quality results.

Conclusion

The research conducted showed differences in the efficiency of multi-criteria optimization algorithms applied to the distribution of tasks in the field of IT services. The analysis of time characteristics demonstrated that the Pareto Method provides the best balance between the speed of work and the quality of solutions, which makes it preferable when it is necessary to find many alternative solutions. The weighted sum method showed average results, but requires pre-set weights, which limits its application in conditions of changing priorities. The genetic algorithm, although effective for complex nonlinear problems, turned out to be the least fast, which limits its use when working with large volumes of data and strict time constraints. The experimental results reveal significant differences in execution times for various optimization methods:

- Genetic Algorithm exhibited the longest execution time, at 0.013633 seconds. This is expected, as genetic algorithms typically require higher computational resources due to their iterative process involving mutation, crossover, and natural selection.

- Pareto Method was the fastest, with an execution time of only 0.000194 seconds. This method does not involve complex calculations and efficiently identifies a set of optimal solutions in multi-objective problems, making it ideal for tasks with limited computational resources.

- Weighted Sum Method had an intermediate execution time of 0.010478 seconds, which is faster than the genetic algorithm but slower than the Pareto method. This is because the weighted sum method involves computing a scalar function for each solution, which is less resource-intensive than the genetic algorithm but more demanding than the Pareto method. Thus, the choice of optimization method depends

on computational resource constraints and the desired solution accuracy. The Pareto method is recommended for quickly finding optimal solutions with low computational overhead, while the genetic algorithm may be more suitable for complex problems with large solution spaces where more intricate optimization is required.

REFERENCES

Alenezi M., Alshammari R., Alrashed R. (2022). Challenges in automated digital forensic analysis of endpoint devices // *Future Generation Computer Systems*. — 2022. — Vol. 128. — Pp. 360–372.

Bai J., Zhang W., Xu H. (2020). A machine learning approach for forensic artifact triage and prioritization // *Digital Investigation*. — 2020. — Vol. 34. — P. 301048.

Bezkorovainyi V., Bezuhla H., Cholombytko D. (2022). Mathematical models of the cyclic work package distribution task // *Press of the Kharkiv National University of Radioelectronics eBooks*. — 2022. — Pp. 7–15. — DOI: 10.30837/MMP.2022.007.

Burkart N., Huber M.F. (2021). A survey on the explainability of supervised machine learning // *Journal of Artificial Intelligence Research*. — 2021. — Vol. 70. — Pp. 245–317.

Callegati F., Cerroni W. (2022). Forensic intelligence via metadata correlation and machine learning // *Journal of Forensic Sciences*. — 2022. — Vol. 67(4). — Pp. 1223–1234.

Eremina I., Lysanov D., Ishmuradova I. I., Isavnin A. (2021). Optimization mathematical model of the distribution of tasks and labor resources in the enterprise // *SHS Web of Conferences*. — 2021. — Vol. 93. — P. 03001. — DOI: 10.1051/SHSCONF/20219303001.

Grod I., Balyk N., Vasylenko Y., Martyniuk S., Oleksiuk V., Barna O. (2022). Web service of works planning using network graph // *Physico-Mathematical Education*. — 2022. — 34(2). — Pp. 18–24. — DOI: 10.31110/2413-1571-2022-034-2-003.

Guo F., Chen L., Li M. (2022). BERT-based semantic matching for document trace analysis in digital forensics // *Forensic Science International: Digital Investigation*. — 2022. — Vol. 40. — P. 301305.

Gupta C., Gupta V. (2024). Enhancing bug allocation in software development: a multi-criteria approach using fuzzy logic and evolutionary algorithms // *PeerJ Computer Science*. — 2024. — Vol. 10. — P. e2111. — DOI: 10.7717/peerj-cs.2111.

Gupta S., Singh R. S. (2024). User-defined weight based multi-objective task scheduling in cloud using whale optimization algorithm // *Simulation Modelling Practice and Theory*. — 2024. — Vol. 133. — P. 102915. — DOI: 10.1016/j.simpat.2024.102915.

Hao Y., Zhao C., Li Z., Si B., Unger H. (2024). A learning and evolution-based intelligence algorithm for multi-objective heterogeneous cloud scheduling optimization // *Knowledge-Based Systems*. — 2024. — DOI: 10.1016/j.knosys.2024.111366.

Hemanth S. V. et al. (2024). Multi-objective Ant Colony Optimization Technique for Task Scheduling in Cloud Computing // *Proceedings of the 2024 International Conference on Artificial Intelligence and Computer Vision*. — 2024. — DOI: 10.1109/icaaic60222.2024.10575423.

Hu Z., Liu W., Ling S., Fan K. (2021). Research on multi-objective optimal scheduling considering the balance of labor workload distribution // *PLOS ONE*. — 2021. — 16(8). — Pp. 1–15. — DOI: 10.1371/journal.pone.0255737.

Ivchenko I. Y., Akhmedova S. A., Aliyeva N. A. (2024). Modeling optimization tasks in IT project management // *Elektrotekhnichni ta Komp'uterni Systemy*. — 2024. — (40). — Pp. 25–36. — DOI: 10.15276/eltecs.40.116.2024.3.

Jafarova Sh. M., Akhmedova S., Aliyeva A. (2024). Research of methods of modeling of mass service enterprise // *Herald of Dagestan State Technical University. Technical Sciences*. — 2024. — 51(3). — Pp. 54–59. — DOI: 10.21822/2073-6185-2024-51-3-54-59.

Jain D., Al-Shammari F., Qureshi B. (2023). Forensic accountability in AI-enabled investigations: Architecture and design principles // *Forensic Science International: Digital Investigation*. — 2023. — Vol. 46. — P. 301510.

Kim Y., Patel R. (2023). Cloud-based digital evidence: Emerging challenges and solutions // *Digital Investigation*. — 2023. — Vol. 44. — P. 301256.

Kolisch R., Heimerl C. (2012). An efficient metaheuristic for integrated scheduling and staffing IT projects based on a generalized minimum cost flow network // *Naval Research Logistics*. — 2012. — 59(2). — Pp. 111–127. — DOI: 10.1002/NAV.21476.

Lamersdorf A., Münch J., Rombach D. (2008). Towards a Multi-criteria Development Distribution Model:

An Analysis of Existing Task Distribution Approaches // *2008 IEEE International Conference on Global Software Engineering*. — 2008. — Pp. 109–118. — DOI: 10.1109/ICGSE.2008.15.

Lavrov E. et al. (2016). Ergonomics of IT outsourcing: Development of a mathematical model to distribute functions among operators // *Eastern-European Journal of Enterprise Technologies*. — 2016. — 2. — Pp. 32–42. — DOI: 10.15587/1729-4061.2016.66021.

Mac Giolla Chriost D., O Ciardhuain S. (2023). Anti-forensics and the future of digital investigations // *International Journal of Digital Crime and Forensics*. — 2023. — 15(1). — Pp. 1–15.

Mannino M., Chen L. (2021). Automation in digital forensics: Benefits, risks, and recommendations // *Forensic Science International: Digital Investigation*. — 2021. — Vol. 37. — P. 301208.

Marturana F., Tacconi S., Me G. (2020). Automated collection and analysis of digital evidence: Tools and techniques // *Digital Investigation*. — 2020. — Vol. 34. — P. 301047.

Mazalov A. N. et al. (2020). Mathematical model for optimizing distributed information systems // *Journal of Physics: Conference Series*. — 2020. — 1679(2). — Pp. 1–7. — DOI: 10.1088/1742-6596/1679/2/022100.

Mazza A., Chicco G., Russo A. (2012). Multi-objective optimization of distribution systems assisted by decision theory criteria // *MEDPOWER 2012*. — 2012. — Pp. 1–6. — DOI: 10.1049/cp.2012.2020.

Narayan R., Liu Y. (2020). Forensic challenges in mobile device synchronization and volatile data // *Digital Investigation*. — 2020. — Vol. 34. — P. 100812.

Navanesan S. et al. (2024). Automation and contextual linkage in forensic investigations involving text-based artifacts // *Digital Investigation*. — 2024. — Vol. 48. — P. 301580.

Nikolaev V., Saenko I. (2024). Optimization of information resources distribution in common information space // *Trudy Uchebnykh Zavedeniy Svyazi*. — 2024. — 10(3). — Pp. 87–103. — DOI: 10.31854/1813-324x-2024-10-3-87-103.

Novozhylova M., Karpenko M. (2024). Solution of a multicriteria assignment problem using a categorical efficiency criterion // *Radio Electronics, Computer Science, Control*. — 2024. — 4. — Pp. 75–84. — DOI: 10.15588/1607-3274-2024-4-7.

Samek W., Müller K.-R., Montavon G. (2021). Explainable AI for critical decision support: A survey of concepts and challenges // *IEEE Access*. — 2021. — Vol. 9. — Pp. 45715–45745.

Schneider L., Weber J., Becker C. (2020). Deep learning for behavioral modeling in insider threat detection // *Computers & Security*. — 2020. — Vol. 92. — P. 101748.

Sha Z., Liang Z., Du X. (2022). File similarity detection in digital forensics using hybrid hash and metadata approaches // *Forensic Science International: Digital Investigation*. — 2022. — Vol. 41. — P. 301338.

Stelly J., Kruse W. (2020). Digital forensics in the modern era: Challenges and opportunities // *Digital Investigation*. — 2020. — Vol. 33. — P. 200901.

Turlakova S. (2022). Research of mathematical methods and models of long-term industrial development // *Economy of Industry*. — 2022. — 4(100). — Pp.53–77. — DOI: 10.15407/econindustry2022.04.053.

Wang J. et al. (2022). Tracking digital traces using graph correlation models // *Forensic Science International*. — 2022. — Vol. 340. — P. 111445.

Xia B., Li C., Zhou Z., Liu J. (2023). Research on Deployment Method of Service Function Chain based on Network Function Virtualization in Distribution Communication Network // *ITNEC 2023*. — 2023. — Pp.1410–1414. — DOI: 10.1109/ITNEC56291.2023.10082364.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 190–207
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.011

УДК 520.6.05

ANALYSIS OF STRAYLIGHT IN CATADIOPTRIC OPTICAL SYSTEMS

B.R. Zhumazhanov[1], A.T. Zhetpisbayeva[2], A.Ye. Kulakayeva[2*], B.S. Zhumazhanov[3]
[1]«Ghalam» LLP, Astana, Kazakhstan;
[2]International Information Technology University, Almaty, Kazakhstan;
[3]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
E-mail: a.kulakayeva@iitu.edu.kz

Berik Zhumazhanov — Head of the Payload and R&D Department, «Ghalam» LLP., 2nd year doctoral student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
http://orcid.org/0000-0001-5926-9619;
Ainur Zhetpisbayeva — associate professor, the Department of Radio Engineering, Electronics and Telecommunications, L.N. Gumilyov Eurasian National University, researcher, the Payload and R&D Department, «Ghalam» LLP, Astana, Kazakhstan
https://orcid.org/0000-0002-4525-5299;
Aigul Kulakayeva — associate professor, the Department of Radio Engineering, Electronics and Telecommunications, International Information Technology University, Almaty, Kazakhstan
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Beksultan Zhumazhanov — design engineer, the Payload and R&D Department, «Ghalam» LLP, Astana, Kazakhstan
https://orcid.org/0009-0000-9493-7491.

Abstract. The article discusses methods for calculating and analyzing unwanted components of the light flux that hit the detector of a catadioptric optical system designed for Earth remote sensing (ERS). With increasing requirements for the accuracy and resolution of observations and the tendency for space technology to be compact, restrictions on the number and size of elements blocking straylight are increasing, given that even a small proportion of parasitic radiation can greatly distort the useful signal, worsening the contrast and reducing the SNR. The objective of the work is to develop and test a comprehensive approach to calculating the amount of parasitic radiation hitting the detector and developing methods for minimizing it. The practical importance of the work is in the fact that the detailed calculated indicators

allow designers to identify sources of straylight and make changes to the system design to minimize it. In the payloads of ERS satellites, minimizing starylight increases the contrast and accuracy of data, ensuring the receipt of clear and informative images, which is necessary for reliable analysis of the state of the environment, agriculture and other ERS tasks.

Keywords: straylight, signal to noise ratio, baffle, remote sensing of the Earth, modeling in Ansys Zemax OpticStudio, catadioptric optical system

For citation: B.R. Zhumazhanov, A.T. Zhetpisbayeva, A.Ye. Kulakayeva*, B.S. Zhumazhanov. Analysis of straylight in catadioptric optical systems// International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 190–207. (In Rus.). https://doi.org/10.54309/IJICT.2025.24.4.011.

Conflict of interest: The authors declare that there is no conflict of interest.

КАТАДИОПТРИКАЛЫҚ ОПТИКАЛЫҚ ЖҮЙЕЛЕРДЕГІ ПАРАЗИТТІК СӘУЛЕЛЕНУДІ ТАЛДАУ

Б.Р. Жумажанов[1], А.Т. Жетписбаева[2], А.Е. Кулакаева[2*], Б.С. Жумажанов[3]

[1]«Ghalam» ЖШС, Астана, Қазақстан;
[2]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[3]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан.
E-mail: a.kulakayeva@iitu.edu.kz

Жумажанов Берик — «Ghalam» ЖШС, пайдалы жүктеме және ғылыми әзірлемелер бөлімінің басшысы. Л.Н.Гумилев атындағы Еуразия ұлттық университетінің 2 курс докторанты, Астана, Қазақстан
http://orcid.org/0000-0001-5926-9619;
Жетписбаева Айнур — «Радиотехника, электроника және телекоммуникациялар» кафедрасының қауымдастырылған профессоры, Л.Н.Гумилев атындағы Еуразия ұлттық университеті, «Ghalam» ЖШС, пайдалы жүктеме және ғылыми әзірлемелер бөлімінің ғылыми қызметкері, Астана, Қазақстан
https://orcid.org/0000-0002-4525-5299;
Кулакаева Айгуль — «Радиотехника, электроника және телекоммуникациялар» **к**афедрасының қауымдастырылған профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Жумажанов Бексултан — «Ghalam» ЖШС, Пайдалы жүктеме және ғылыми әзірлемелер бөлімінің Инженер-конструкторы, Астана, Қазақстан https://orcid.org/0009-0000-9493-7491.

Аннотация. Мақалада Жерді қашықтықтан зондтауға (ЖҚЗ) арналған

катадиоптриялық оптикалық жүйенің детекторына түсетін жарық ағынының паразиттік компоненттерін есептеу және талдау әдістері қарастырылады. Бақылаулардың дәлдігі мен ажыратымдылық қабілетіне қойылатын талаптар артқан сайын, ал ғарыштық техниканың ықшамдалу үрдісін ескере отырып, паразиттік сәулеленуді бөгейтін элементтердің саны мен өлшемдеріне қойылатын шектеулер де күшейе түсуде. Бұл ретте тіпті аз ғана мөлшердегі паразиттік сәулеленудің өзі пайдалы сигналды едәуір бұрмалап, контрастты төмендетіп және С/Ш қатынасын нашарлатуы мүмкін. Жұмыста детекторға түсетін паразиттік сәулелену мөлшерін есептеудің кешенді тәсілін әзірлеу және оны азайту әдістерін ұсыну міндеті қойылған. Жұмыстың практикалық маңызы болып есептелген көрсеткіштер жобалаушыларға паразиттік сәулелену көздерін анықтауға және оны азайту үшін жүйе құрылымына өзгерістер енгізуге мүмкіндік береді. ЖҚЗ ғарыш аппараттарының пайдалы жүктемесінде паразиттік сәулеленуді азайту контраст пен деректердің дәлдігін арттырады, бұл өз кезегінде қоршаған ортаның жай-күйін, ауыл шаруашылығын және ЖҚЗ-ға қатысты басқа да міндеттерді сенімді талдауды қамтамасыз етеді.

Түйін сөздер: паразиттік сәулелер, сигнал / шу қатынасы, бленда, Жерді қашықтықтан зондтау, Ansys Zemax OpticStudio модельдеу, катадиоптрикалық оптикалық жүйе

Дәйексөздер үшін: Жумажанов Б.Р., Жетписбаева А.Т., Кулакаева А.Е., Жумажанов Б.С. катадиоптикалық оптикалық жүйелердегі паразиттік сәулеленуді талдау////Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 23. Стр. 190–207. (орыс тіл.). https://doi. org/10.54309/IJICT.2025.24.4.011.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

АНАЛИЗ ПАРАЗИТНОГО ИЗЛУЧЕНИЯ В КАТАДИОПТРИЧЕСКИХ ОПТИЧЕСКИХ СИСТЕМАХ

Б.Р. Жумажанов[1], А.Т. Жетписбаева[2], А.Е. Кулакаева[2*], Б.С. Жумажанов[3]

[1]ТОО «Ghalam», Астана, Казахстан;

[2]Международный университет информационных технологий, Алматы, Казахстан;

3Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан.

E-mail: a.kulakayeva@iitu.edu.kz

Жумажанов Берик — руководитель отдела полезной нагрузки и научных разработок, ТОО «Ghalam». докторант 2 года обучения, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан http://orcid.org/0000-0001-5926-9619;

Жетписбаева Айнур — ассоциированный профессор, кафедра радиотехники, электроники и телекоммуникации, Евразийский национальный университет имени Л.Н. Гумилева, научный сотрудник отдела полезной нагрузки и научных разработок, ТОО «Ghalam», Астана, Казахстан
https://orcid.org/0000-0002-4525-5299;
Кулакаева Айгуль — ассоциированный профессор, кафедра радиотехники, электроники и телекоммуникации, Международный университет информационных технологий, Алматы, Казахстан
E-mail: a.kulakayeva@iitu.edu.kz, https://orcid.org/0000-0002-0143-085X;
Жумажанов Бексултан — инженер-конструктор, отдел полезной нагрузки и научных разработок, ТОО «Ghalam», Астана, Казахстан
https://orcid.org/0009-0000-9493-7491.

Аннотация. В статье рассматриваются методы расчета и анализа нежелательных компонентов светового потока, которые попадают на детектор катадиоптрической оптической системы, предназначенной для дистанционного зондирования Земли (ДЗЗ). С ростом требований к точности и разрешающей способности наблюдений при тенденции к компактности космической техники, возрастают ограничения к количеству и размерам элементов блокирования паразитного излучения, при том, что даже незначительная доля паразитного излучения способна сильно исказить полезный сигнал, ухудшая контраст и снижая отношение С/Ш. В работе ставится задача разработать и опробовать комплексный подход расчета количества паразитного излучения, попадающего на детектор и разработки методов их минимизации. Практическая важность работы заключается в том, что детально рассчитанные показатели позволяют проектировщикам выявить источники паразитного излучения и внести изменения в конструкцию системы для его минимизации. В полезных нагрузках космических аппаратов ДЗЗ минимизация паразитного излучения повышает контрастность и точность данных, обеспечивая получение четких и информативных изображений, что необходимо для достоверного анализа состояния окружающей среды, сельского хозяйства и других задач ДЗЗ.

Ключевые слова: паразитное излучение, отношение сигнал/шум, бленда, дистанционное зондирование Земли, моделирование в Ansys Zemax OpticStudio, оптическая система

Для цитирования: Жумажанов Б.Р., Жетписбаева А.Т., Кулакаева А.Е., Жумажанов Б.С. Анализ паразитного излучения в катадиоптрических оптических системах ////Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 190–207. (На рус.). https://doi. org/10.54309/IJICT.2025.24.4.011.

Конфликт интересов: авторы заявляют об отсутствии конфликта

интересов.

Данное исследование финансировалось Аэрокосмическим комитетом Министерства цифрового развития, инноваций и аэрокосмической промышленности Республики Казахстан (BR21982462 «Разработка и тестирование малого спутника на основе отечественных технологий»).

Введение

Паразитное излучение, проникающее на детектор помимо основного оптического пути, способно существенно ухудшить качество получаемых данных. В высокоточных оптических системах, таких как астрономические телескопы или аппараты для дистанционного зондирования Земли (ДЗЗ), даже незначительная доля рассеянного или прямого света может искажать результаты измерений и затруднять анализ слабых сигналов. По мере увеличения требований к точности наблюдений, включая задачи мониторинга поверхности и атмосферы Земли, актуальной становится проблема выявления возможных источников паразитного излучения в оптическом тракте и разработки эффективных методов его подавления.

В сложных конструкциях оптических систем, где помимо основных зеркал или линз могут присутствовать дополнительные оптические и механические элементы (бленды, оправы, внутренние диафрагмы и т. д.), формируется множество каналов, через которые избыточное излучение может достичь детектора. Это явление приводит к снижению контрастности изображения, увеличению уровня шума и затрудняет выполнение научных и прикладных задач. Для приборов, осуществляющих наблюдения Земли, минимизация паразитного излучения особенно критична, так как оно может маскировать слабые сигналы от земной поверхности, снижая точность и достоверность получаемой информации.

В настоящее время представлены комплексные оптикомеханические решения для подавления паразитного света в самых разных системах. Так, для фотоэлектрической системы зондирования была разработана асферическая воздушная камера с блендой и светозащитными кольцами, в которой моделирование в TracePro показало резкое снижение значения коэффициента пропускания точечного источника уже при угле падения > 0,5°, а к 30° – до ~10^{-8}, что значительно улучшает контраст изображений (Li et al., 2022: 10). Аналогично, в подводном спектральном радиометре с плоскополевым вогнутым голографическим дифракционным элементом проанализированы источники паразитного света и предложены методы его подавления. Благодаря этому удалось снизить уровень засветки до ~10^{-4}, повысить отношение сигнал/шум и подтвердить спектральное разрешение < 3 нм, а также установить количественную зависимость между спектральной освещенностью и цифровым выходом прибора (Zhang et al., 2024: 1). Для космических оптикоэлектронных систем был испытан бронеситалл – легкая и технологичная

бронестеклокерамика – вместе с сегментированными зеркалами, удлиненными блендами и переходными блоками. Баллистические испытания показали, что единичные каверны на кварце лишь незначительно влияют на оптику, тогда как алюминиевые зеркала деформируются и дают рикошет, подтверждая эффективность предложенных пассивных средств защиты (Медунецкий и др. 2023: 587-591). В компактной системе SOCD с «наводящим» зеркалом и афокальным участком авторы учли ошибки формы и микрошероховатости, спроектировали фигурную входную диафрагму и нестандартный козырек, оптимизировали крепления и покрытия, что позволило в моделировании и экспериментах сократить засветку при углах до 60–80° и повысить точность независимо от ориентации платформы (Han et al., 2023: 2–8). При моделировании рамановского спектрометра Черни–Тернера с алгоритмом POP в Ansys Zemax OpticStudio были проанализированы spotдиаграммы, MTF и encircled energy в диапазоне 530–630 нм (возбуждение 527 нм), что показало высокую эффективность формирования спектра и достаточное пространственное разрешение для компактного недорогого прибора (Naeem et al., 2022: 2–6). Для средневолнового ИКФурьеспектрометра на двойном качающемся зеркале предложили наружный кожух с углом избегания $\geq 30°$ и внутренние кольцевые диафрагмы на зеркалах Кассегрена, а также введение клинового угла и наклона у пластин, что снизило PST с 10^{-4} до $10^{-7}…10^{-8}$ и восстановило контраст интерференционных полос (Ben et al., 2024: 6–15). Также, на основе моделирования орбитальной геометрии спутника FY-3C в STK и трассировки лучей в TracePro с учетом BRDF-параметров конструктивных поверхностей прибора VIRR показано, что внешнее паразитное излучение от Солнца наиболее интенсивно при входе и выходе из тени Земли (углы солнечного зенита 60–90°). Сравнение моделируемой картины засветки с реальными артефактами на длине волны 3,7 μм (Band 3) подтвердило достоверность метода, который может быть адаптирован для анализа других орбитальных платформ и приборов (Zhao et al., 2021: 3–16). В интегрированной VIS–IR–THzсистеме изза роста отражения в ТГц диапазоне удлинили переднюю бленду до длины всего оптического пути с кольцевыми диафрагмамивставками, установили экран между первым и третьим зеркалами с наклонными стенками и отражающую диафрагму перед фокальной плоскостью, а поверхности «осветлены» и охлаждены до 150–200 K, что позволило снизить PST в VIS–IR до $10^{-5}…10^{-7}$ при 10–20° и в ТГц (343 ГГц) до $10^{-7}…10^{-8}$, а собственное излучение системы стало приемлемым (Jiang et al., 2–13). Для спектрометра FY3D/MERSI на солнечносинхронной орбите определили критические углы солнечного освещения при заходе/выходе из тени, спроектировали sunshade и inlet hood, добавили внутренние элементы - диафрагму Лио, ограничитель поля зрения, локальные бленды и специальные покрытия, благодаря чему PST в каналах VIS–NIR упал до $10^{-4}…10^{-5}$ при 10–20°, что подтвердили лабораторные измерения, обеспечив паразитную засветку < 1 % от полезного сигнала и создав типовую методику для широкоуголь-

ных спектральных приборов на метеоспутниках (Chen et al., 2–16). В целом, установка для PSTизмерений с полихроматом и пятью лазерами (λ 0,447–2,2 μм), зеркальным коллиматором Ø 1 м, двухколонной черной камерой и калибровочной линзой f = 250 мм, F/5, поле ±6° продемонстрировала диапазон PST 10^{-3}…10^{-10} при θ = ±5…±60° (Lu et al., 33939-33946).

Основная цель данного исследования состоит в разработке и проверке методики расчета паразитного излучения в катадиоптрической системе на базе программного комплекса Ansys Zemax OpticStudio, а также в оценке его влияния на качество получаемого изображения. Для достижения указанной цели проведен комплекс симуляций, охватывающих как оптические, так и системные аспекты.

На первом этапе выполнено моделирование распространения света с использованием трассировки лучей в программном комплексе Ansys Zemax OpticStudio. Это позволило идентифицировать основные источники паразитного излучения, такие как рассеяние на оптических элементах и отражения от внутренних поверхностей системы. Для более детального анализа проведено моделирование рассеяния света на оптических поверхностях с учетом шероховатостей, дефектов и загрязнений, а также моделирование многократных отражений внутри оптической системы, включая влияние антибликовых покрытий и диафрагм.

Далее проведено моделирование взаимодействия света с механическими элементами конструкции, такими как бленды, оправы и крепления. Это позволило оценить их вклад в формирование паразитного излучения. Кроме того, выполнены тепловые и механические симуляции для анализа влияния деформаций оптической системы, вызванных температурными и механическими нагрузками, на траектории лучей и уровень паразитного излучения.

Для оценки влияния паразитного излучения на качество изображения проведено моделирование работы детектора, включая анализ уровня шума и динамического диапазона. Также выполнены симуляции формирования изображения с учетом паразитного излучения, что позволило оценить его влияние на контрастность, разрешение и общее качество изображения.

На системном уровне проведено моделирование полной оптической системы с учетом всех элементов (оптика, механика, детектор) для комплексной оценки влияния паразитного излучения. Дополнительно выполнены симуляции в условиях эксплуатации, включая влияние внешних факторов, таких как солнечное излучение и фоновая засветка, что особенно актуально для космических аппаратов дистанционного зондирования Земли.

Для разработки методов подавления паразитного излучения проведена оптимизация конструкции оптической системы с использованием различных конфигураций. Также протестированы защитные меры, такие как бленды, диафрагмы и антибликовые покрытия, для оценки их эффективности.

На заключительном этапе выполнены алгоритмические симуляции для

разработки и тестирования методов обработки изображений, направленных на подавление паразитного излучения. Это включает методы вычитания фона, фильтрации шумов и калибровки системы для компенсации влияния паразитного излучения.

Проведенные симуляции позволили всесторонне изучить влияние паразитного излучения на качество изображения и разработать эффективные методы его минимизации. Полученные результаты могут быть использованы для оптимизации конструкции катадиоптрических систем и повышения точности получаемых данных в условиях, где паразитное излучение может существенно влиять на результаты измерений.

Совместное использование в данной оптической системе наружной и внутренней бленд в сочетании с антиотражательными покрытиями обеспечивает наиболее выраженный эффект снижения влияния паразитного излучения. Такой подход не допускает прямого попадания света на детектор и за счет многократного отражения на внутренних элементах оптической системы минимизирует уровень нежелательного светового фона.

Новизна данного исследования заключается в интегральном подходе, который объединяет моделирование различных типов избыточных компонентов светового потока – от рассеянного фона под малыми углами до мощного прямого попадания – с анализом разнообразных конфигураций использования бленд и покрытий в оптической системе. Использование программного комплекса Ansys Zemax OpticStudio позволяет не только оценивать количественные показатели, такие как суммарная мощность и количество паразитных лучей, достигших детектора, но и детально анализировать траектории их распространения, идентифицировать основные источники нежелательного излучения и оптимизировать конструкцию системы для минимизации его влияния. Кроме того, с помощью Ansys Zemax OpticStudio возможно моделировать различные сценарии эксплуатации, включая изменение углов падения света, уровней засветки и температурных условий, что обеспечивает всестороннюю оценку устойчивости оптической системы к паразитному излучению. Это особенно важно для разработки высокоточных приборов, работающих в сложных условиях, таких как космические аппараты дистанционного зондирования Земли, где даже незначительные искажения могут существенно повлиять на качество получаемых данных.

Результаты проведенного моделирования могут быть применены при проектировании высокоточных оптических приборов, включая системы, работающие на орбите и предназначенные для дистанционного зондирования Земли. В таких приложениях надежная изоляция детектора от паразитного излучения является критически важной для обеспечения высокой точности и достоверности получаемых данных.

Материалы и методы

Объектом исследования выступает зекально-линзовая оптическая

система, разработанная специалистами ТОО «Ghalam», предназначенная для использования в составе полезных нагрузок космических аппаратов дистанционного зондирования Земли (КА ДЗЗ). Для анализа паразитного излучения и его влияния на качество изображения использовался программный комплекс Ansys Zemax OpticStudio, который поддерживает моделирование распространения света в режиме непоследовательного распространения излучения (Non-sequential mode), что позволяет учитывать многократные отражения, рассеяние на шероховатых поверхностях, а также взаимодействие света с трехмерными механическими элементами, такими как бленды, оправы и диафрагмы. Этот режим также обеспечивает возможность анализа случайных путей проникновения света, включая отражения и рассеяние, что критически важно для изучения паразитного излучения и его влияния на качество изображения. Параметры оптического телескопа приведены в таблице 1 (В симуляции все покрытия близки к идеальному).

Таблица 1. Основные параметры телескопа Ричи–Кретьен, использованного в моделировании паразитного излучения

Параметры	Значения
Тип телескопа	Ричи-Кретьен
Фокусное расстояние	550 мм
Длина волны	Видимый свет 400-700 нм
Апертура	85 мм
Относительное отверстие	1/6,47
Количество лучей	Для сценария ambient straylight использовался один источник с количеством лучей 10^6, а для сценария direct-hit straylight – несколько источников с суммарным количеством до 10^5
Покрытие линз	AR покрытия с идеальной пропускной способностью
Зеркала	С идеальным отражательным покрытием
Бленды и поверхность трубы	Полностью поглощают попадающий свет
Размер детектора	4096×3072 пикселей, однако для ускорения моделирования он был уменьшен до 250×250 пикселей

Алгоритм проведения моделирования паразитного излучения в исследуемой оптической системе приведен на рисунке 1.

Основные параметры, анализируемые в ходе моделирования, включали суммарную мощность паразитного излучения, максимальную освещенность на детекторе и количество лучей, достигающих детектора в обход основного оптического пути.

Экспериментальная часть исследования заключается в проведении виртуальных симуляций, включающих моделирование многократных отражений, рассеяния света и взаимодействия с механическими элементами, такими как бленды и диафрагмы.

Рис. 1. Алгоритм моделирования паразитного излучения

Анализируются различные конфигурации системы, включая антиотражательные покрытия, для оценки их эффективности в минимизации паразитного излучения. Также моделируются условия эксплуатации, такие как изменение углов падения света и уровней засветки, что позволяет выявить оптимальные конструктивные решения для повышения качества изображения. Все симуляции проводились на этапе проектирования, что позволило принять меры для снижения уровня паразитного излучения.

Методология исследования базируется на сравнительном анализе различных конфигураций оптической системы, включая варианты с использованием наружных и внутренних бленд, а также антиотражательных покрытий. Такой подход позволяет выявить оптимальный набор защитных мер, обеспечивающих минимальное проникновение нежелательных лучей к детектору.

Полученные результаты могут быть применены при проектировании высокоточных оптических приборов для дистанционного зондирования Земли, где минимизация паразитного излучения является критически важной для обеспечения достоверности и точности получаемых данных.

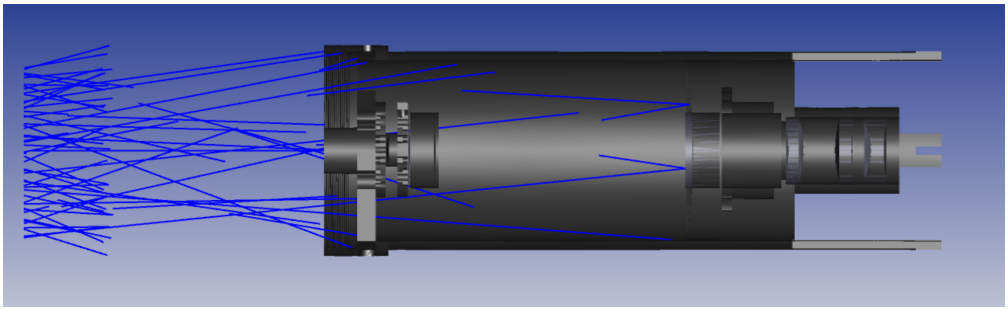
Результаты и обсуждение

Анализ паразитного излучения необходим для минимизации нежелательных помех, которые могут исказить данные, получаемые приборами дистанционного зондирования Земли. Данный анализ позволяет выявить источники паразитного излучения, такие как внутренние отражения и внешние факторы, а также определить меры по их снижению. В результате повышается отношение сигнал/шум и возрастает точность регистрируемых данных.

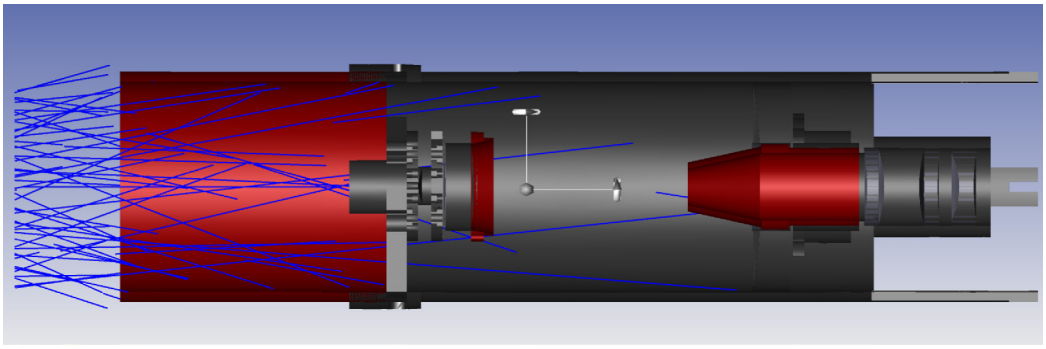
Паразитное излучение окружающей среды (Ambient straylight) – рассеянный свет, поступающий в оптическую систему из внешних источников, не связанных напрямую с целевым объектом наблюдения. Такой свет может проникать в оптическую систему, вызывая искажения изображения и снижение

его контраста, что ухудшает общую производительность системы.

Для оценки воздействия паразитного излучения окружающей среды в программном комплексе Ansys Zemax OpticStudio перед 3D-моделью оптической системы устанавливается источник света, имитирующий данное излучение. На рисунке 2 представлена 3D-модель с источником света и трассировкой 100 лучей. В ходе анализа было проверено 10^6 лучей от 20 источников излучения с углами падения на телескоп от 1,5° до 30°. Рисунок 2 (а) демонстрирует телескоп без бленд, тогда как вариант (б) иллюстрирует ту же конструкцию, но с добавлением всех необходимых экранов (красным цветом выделены наружная бленда, а также бленды для вспомогательного и главного зеркал). В каждом случае прослеживается траектория лучей, чтобы оценить, насколько защитные элементы предотвращают попадание нежелательного излучения внутрь прибора.



а) система без бленд



б) система с установленными блендами

Рис. 2. Трассировка световых лучей внутри телескопа

Сопоставление рисунков 2(а) и 2(б) подтверждает, что установка бленд значительно уменьшает количество лучей, способных попасть на основные оптические поверхности. Это приводит к сокращению нежелательных компонентов световых потоков и повышению качества регистрируемого изображения. Таким образом, внедрение экранирующих конструкций является одним из основных способов борьбы с нежелательным рассеянным светом в телескопических системах, применяемых в том числе и для дистанционного

зондирования Земли.

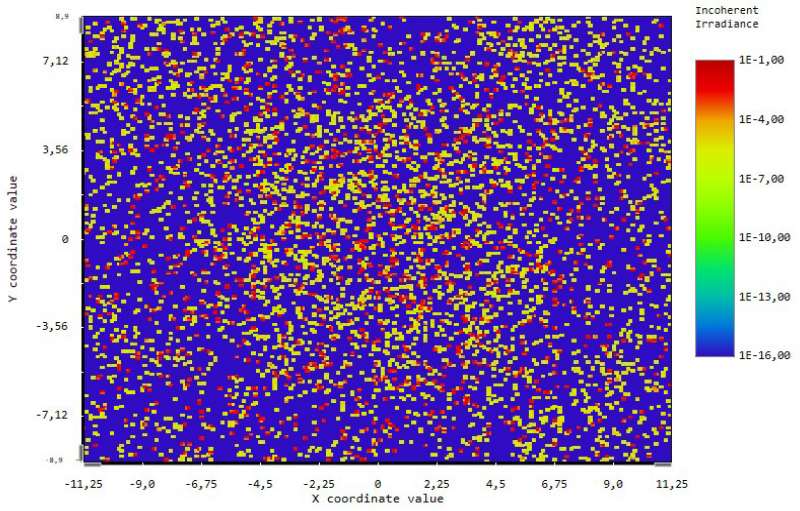
Кроме этого, изображения на детекторах показаны на рисунке 3, а подробные результаты приведены в таблице 2. Результаты моделирования, демонстрирующие распределение нежелательных компонентов световых потоков на детекторе при отсутствии защитных экранов (рисунок 3(а)), а на рисунке 3(б), та же система, но с установленными блендами. В первом случае нежелательные лучи беспрепятственно проникают на детектор и увеличивают общий фоновый уровень, а во втором блокирующие элементы заметно снижают паразитное освещение.

Сопоставление рисунков 3(а) и 3(б) наглядно свидетельствует о высокой эффективности бленд при подавлении нежелательных лучей. Это обеспечивает более низкий уровень шумов и, как следствие, улучшает контрастность изо-бражения, что особенно важно для точных измерений и наблюдений. Прямое попадание рассеянного света (Direct hit straylight) – свет, который проникает в оптическую систему непосредственно от интенсивного внешнего источника, распространяясь без значительного рассеяния. Это приводит к появлению артефактов и снижению качества изображения из-за прямого воздействия на элементы оптики.

Таблица 2. Результаты анализа паразитного излучения прямого попадания

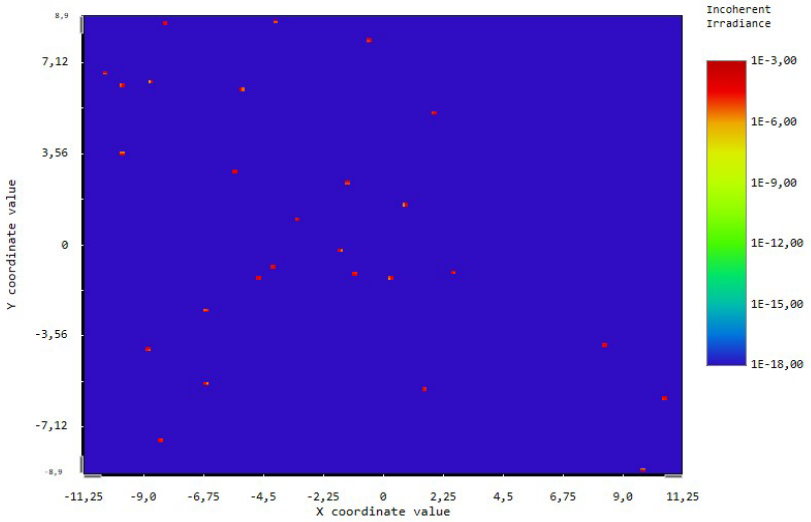
Конфигурация телескопа	Попадание лучей	Общая мощность, Вт	Максимальная освещенность, Вт/см^2
Без бленд, без покрытия	4813	9,15E-04	1,46E-02
Наружная бленда, без покрытия	3622	7,03E-04	1,41E-02
Бленда главного зеркала, без покрытия	500	1,35E-06	3,54E-04
Наружная и ГЗ бленды, без покрытия	401	1,23E-06	3,52E-04
Без бленд, AR покрытие	998	8,14E-04	1,73E-02
Наружная бленда, AR покрытие	27	2,89E-07	1,42E-04
Бленда ГЗ, AR покрытие	994	7,92E-04	1,86E-02
Наружная и ГЗ бленды, AR покрытие	708	5,55E-04	1,43E-02

Для анализа прямого попадания паразитного излучения перед телескопом на большом расстоянии размещаются 8 источников света с углами падения от 5° до 30°. В рамках исследования было проверено 10^5 лучей для каждого источника. На рисунке 4 показана ситуация, когда яркий внешний источник света попадает на телескоп практически напрямую, без значительного рассеяния.



Detector Image: Irradiance	
Ghalam optic system 14.08.2025 Detector 20, NSCG Surface 1: Field 1 Size 22,500 W X 17,800 H Millimeters, Pixels 250 W X 250 H, Total Hits = 4890 Peak Irradiance : 1,5465E-02 Watts/cm^2 Total Power : 9,1639E-04 Watts	Zemax Ansys Zemax OpticStudio 2024 R1.00 Ghalam_MRC_550_VIS.MAF.14.ao-SLAAL.ZMX Configuration 1 of 1

а) изображение на детекторе телескопа без бленд

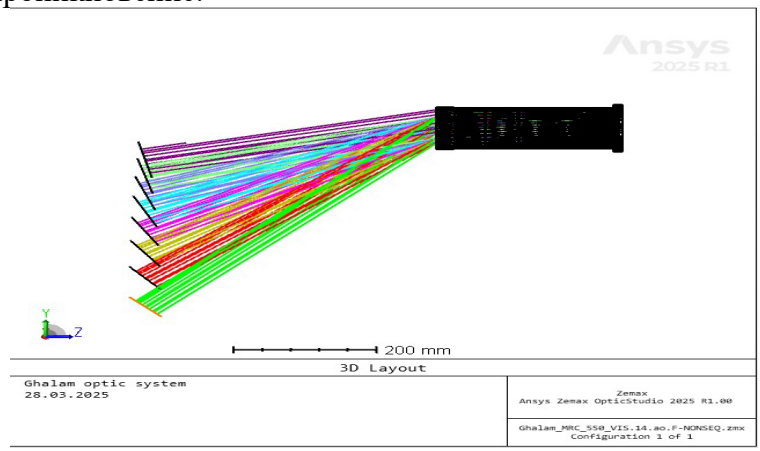


Detector Image: Irradiance	
Ghalam optic system 14.08.2025 Detector 20, NSCG Surface 1: Field 1 Size 22,500 W X 17,800 H Millimeters, Pixels 250 W X 250 H, Total Hits = 27 Peak Irradiance : 1,4047E-04 Watts/cm^2 Total Power : 2,9968E-07 Watts	Zemax Ansys Zemax OpticStudio 2024 R1.00 Ghalam_MRC_550_VIS.MAF.14.ao-SLAAL.ZMX Configuration 1 of 1

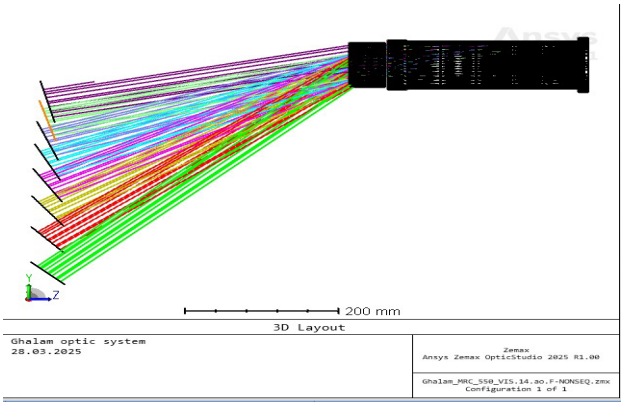
б) изображение на детекторе телескопа с блендами

Рис. 3. Изображение на детекторе телескопа

На рисунке 4(а) представлена конфигурация без бленд, а на рисунке 4(б) показана та же система, но уже с установленными экранами (наружная бленда, а также бленды для вторичного и главного зеркал, выделенные красным цветом). Цветные лучи отражают ход света от источника к оптической системе и позволяют наглядно оценить, насколько эффективно экраны блокируют нежелательное проникновение.



а) в телескопе без бленд

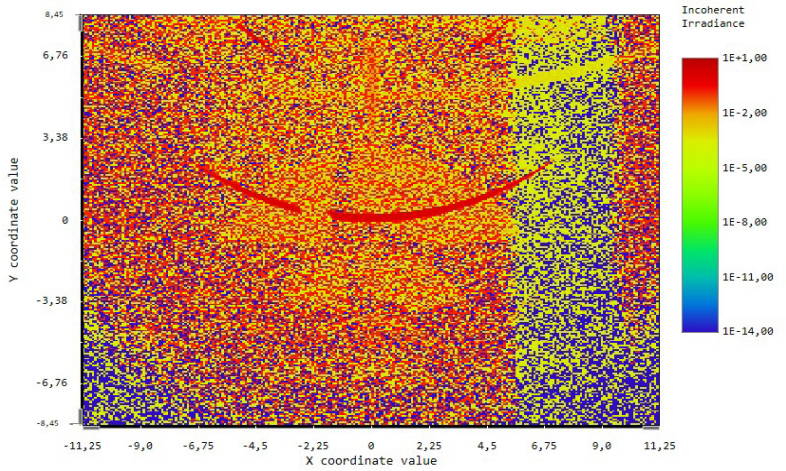


б) в телескопе со всеми блендами

Рис. 4. Ход лучей при симуляции прямого попадания рассеянного света

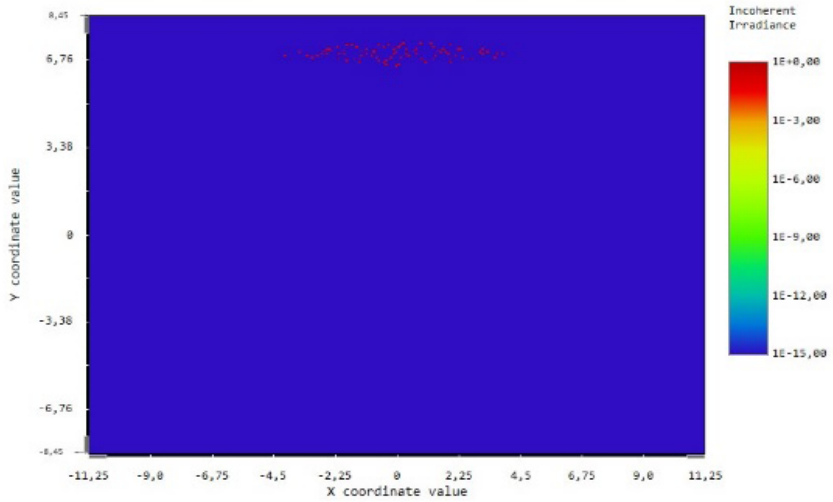
Таким образом, использование защитных экранов становится критически важным при наличии интенсивных источников света, способных вызывать ощутимых нежелательных компонентов светового потока в оптических системах.

На рисунке 5 показаны распределения нежелательного освещения на детекторе при двух различных конфигурациях телескопа. Рисунок 5(а) когда бленды отсутствуют, и рисунок 5(б) когда они установлены в полном объеме. На каждом из изображений цветовая шкала показывает уровень интенсивности, что позволяет оценить общее количество посторонних лучей и характер их распределения по площади детектора.



Detector Image: Incoherent Irradiance	
Ghalam optic system 14.08.2025 Detector 38, NSCG Surface 1: Field 1 Size 22,500 W X 16,900 H Millimeters, Pixels 250 W X 250 H, Total Hits = 124548 Peak Irradiance : 1,7708E+00 Watts/cm^2 Total Power : 2,4166E-01 Watts	Zemax Ansys Zemax OpticStudio 2024 R1.00 Ghalam_MRC_550_VIS.MAF.14.aoSLADH.ZMX Configuration 1 of 1

а) изображение на детекторе телескопа без бленд



Detector Image: Incoherent Irradiance	
Ghalam optic system 14.08.2025 Detector 38, NSCG Surface 1: Field 1 Size 22,500 W X 16,900 H Millimeters, Pixels 250 W X 250 H, Total Hits = 84 Peak Irradiance : 2,4138E-01 Watts/cm^2 Total Power : 6,1157E-04 Watts	Zemax Ansys Zemax OpticStudio 2024 R1.00 Ghalam_MRC_550_VIS.MAF.14.aoSLADH.ZMX Configuration 1 of 1

б) изображение на детекторе телескопа со всеми блендами

Рис.5. Влияние бленд на изображение телескопа

Таблица 3 содержит сравнительные данные по различным вариантам конструкции телескопа в условиях прямого попадания внешнего источника на оптическую систему. Для каждой конфигурации приводятся три основных параметра: общее число лучей, достигших детектора, их суммарная мощность (Вт) и максимальная освещенность детектора (Вт/см^2). Такие показатели позволяют оценить, как конкретное сочетание бленд и антирефлектных покрытий влияет на уровень нежелательного освещения.

Таблица 3. Результаты анализа паразитного излучения прямого попадания

Конфигурация телескопа	Попадание лучей	Общая мощность, Вт	Максимальная освещенность, Вт/см^2
Без бленд, без покрытия	124703	2,41E-01	1,80E+00
Наружная бленда, без покрытия	77482	1,43E-01	1,78E+00
Бленда главного зеркала, без покрытия	16997	1,47E-03	2,04E-01
Наружная и ГЗ бленды, без покрытия	15174	7,11E-04	2,05E-01
Без бленд, AR покрытие	33594	2,79E-01	2,12E+00
Наружная бленда, AR покрытие	20344	1,65E-01	2,10E+00
Бленда ГЗ, AR покрытие	192	1,46E-03	2,39E-01
Наружная и ГЗ бленды, AR покрытие	90	6,12E-04	2,40E-01

Из приведенных сведений видно, что наибольшее число лучей, а также самые высокие мощность и освещенность фиксируются в конфигурациях без бленд и без специальных покрытий. Напротив, комбинированное использование наружной бленды, бленды главного зеркала (ГЗ) и антиотражающее покрытие (AR-покрытие) дает наиболее эффективное подавление паразитного света. Данные результаты подтверждают необходимость комплексного подхода к защите системы, потому что нежелательный или слишком яркий свет может перегрузить чувствительный элемент, исказив полезный сигнал и приводя к неточным результатам.

Полученные результаты наглядно демонстрируют, что основной вклад в общий уровень нежелательного света в телескопической системе вносят прямые пролеты от ярких источников при отсутствии достаточной экранировки, а также

многократные отражения внутри прибора. Установка наружной и внутренней бленд позволила существенно сократить число паразитных лучей и заметно уменьшить общую мощность, регистрируемую детектором.

Ограничением данного исследования является идеализированная среда компьютерного моделирования, в которой не учитываются возможные неточности сборки оптики, загрязнения покрытий и нестабильность внешних источников. Также в реальных условиях могут проявляться дополнительные факторы, связанные с дифракцией на краях бленд или со смещением оптических элементов. Тем не менее, проведенные расчеты дают достоверную оценку основных путей проникновения нежелательного освещения и способов их блокировки, что очень важно при проектировании приборов для дистанционного зондирования Земли.

Практическая значимость данной работы заключается в том, что инженеры и конструкторы, получив готовый набор количественных показателей (число лучей, мощность, максимальная освещенность), могут заранее планировать расстановку бленд и выбор антирефлектных покрытий, не доводя дело до дорогостоящих переделок после изготовления системы.

Однако все еще остаются открытыми вопросы по более детальному учету рассеяния на микрошероховатостях и реальных свойствах материалов, особенно в широком спектральном диапазоне. Также требует внимания моделирование мелких механических деталей, способных вызывать неоднородное рассеяние или непрогнозируемые зеркальные блики.

Заключение

Проведенные моделирования и расчеты четко показали, что при отсутствии эффективных мер экранирования даже небольшое количество нежелательных компонентов светового потока, попадающих в телескоп, может существенным образом искажать изображение. В результате снижается контраст, понижается отношение сигнал/шум и возрастает риск перенасыщения детектора. Установка наружной бленды, дополнительной бленды главного зеркала и применение антирефлектных покрытий позволили на порядок сократить интенсивность и число паразитных лучей, тем самым значительно повысив качество регистрируемых данных. Это особенно важно для аппаратов дистанционного зондирования Земли, где избыточное освещение от посторонних источников способно скрыть слабые сигналы, исказить измерения и усложнить научный анализ. Значимость таких результатов заключается в том, что предложенная методика позволяет проектировщикам еще на стадии виртуального моделирования выявлять основные пути проникновения лишнего света и оптимизировать расстановку экранирующих элементов и покрытий, избегая дорогостоящих корректировок после изготовления приборов.

Также рекомендуется заранее закладывать в конструкцию систему экранирующих элементов, ориентируясь на расчеты в программных пакетах наподобие Ansys Zemax OpticStudio, с учетом возможного внешнего рассеянного

света и прямых попаданий ярких источников. Качественное просветление оптики и обеспечение надежного блокирования самых проблемных путей проникновения лучей позволяют избежать существенного увеличения уровня фоновой засветки и гарантировать высокую точность получаемых данных.

REFERENCES

Ben C., Shen H., Yu X., Meng L., Cheng H., Jia P. (2024). Stray light analysis and suppression for an infrared Fourier imaging spectrometer // Photonics. — 2024. — Vol. 11. — No. 2. — P. 173.

Chen S., Niu X. (2023). Analysis and suppression design of stray light pollution in a spectral imager loaded on a polar-orbiting satellite // Sensors. — 2023. — Vol. 23. — No. 17. — P. 7625.

Han P., Guo J., Zhang M., Xue R., Ren G., Liu Y. (2023). Stray light control for the space-agile optical system with a pointing mirror // Applied Optics. — 2023. — Vol. 62. — No. 12. — Pp. 3132–3141.

Jiang H., Niu X. (2023). Stray light analysis and suppression of the visible to terahertz integrated cloud detection optical system // Sensors. — 2023. — Vol. 23. — No. 8. — P. 4115.

Li J., Yang Y., Qu X., Jiang C. (2022). Stray light analysis and elimination of an optical system based on the structural optimization design of an airborne camera // Applied Sciences. — 2022. — Vol. 12. — No. 4. — P. 1935.

Lu Y., Xu X., Zhang N., Lv Y., Xu L. (2024). Study on stray light testing and suppression techniques for large-field of view multispectral space optical systems // IEEE Access. — 2024. — Vol. 12. — Pp. 33938–33948.

Медунецкий В. М., Добряков Б. Н., Солк С. В., Меркулов Ю. Ю., Сильников Н. М. (2023). Экспериментальные методы и средства защиты оптико-электронных систем космического базирования от механических воздействий // Известия высших учебных заведений. Приборостроение. — 2023. — Т. 66. — № 7. — С. 585–593.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 208–218
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.012

УДК 004.931

DEVELOPMENT OF DOCUMENT ACCOUNTING AND DIGITALIZATION SYSTEMS WITH INTEGRATION OF ARTIFICIAL INTELLIGENCE

A. Kassymova[1], R. Tlegenov[1*] A. Alimagambetova[2], G. Murzabekova[3], M. Kassim[4]

[1]Zhangir Khan West Kazakhstan Agrarian Technical University, Uralsk, Kazakhstan;
[2]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[3]Kazakh Agrotechnical Research University named after S. Seifullin, Astana, Kazakhstan;

[3] MARA University of Technology, Malaysia.
E-mail: rinattlegenov1@gmail.com

Akmaral Kassymova — associate Professor, Candidate of Pedagogical Sciences, Agrarian-Technical University of Western Kazakhstan University named after Zhangir Khan, Uralsk, Kazakhstan
E-mail: kasimova_ah@mail.ru, https://orcid.org/0000-0002-4614-4021;
Rinat Tlegenov — Master's student, West Kazakhstan Agrarian and Technical University named after Zhangir Khan, Uralsk, Kazakhstan
https://orcid.org/0009-0004-5476-5970**;**
Ainagul Alimagambetova — L.N. Gumilyov Eurasian National University, Senior Lecturer of the Department of Information Systems, Candidate of Physical and Mathematical Sciences, Astana, Kazakhstan
https://orcid.org/0000-0002-9859-2029;
Gulden Murzabekova — Kazakh Agrotechnical Research University named after S. Seifullin, Associate Professor of the Department of Computer Science, Candidate of Physical and Mathematical Sciences, Astana, Kazakhstan
https://orcid.org/0000-0001- 9807-5200;
Kassim Murizah — PhD, Associate Professor, MARA University of Technology, Malaysia
https://orcid.org/0000-0002-8494-4783.

Abstract. In this article, the issues of developing and implementing document accounting and digitalization systems using artificial intelligence (AI) technologies

are considered. The relevance of the topic is due to the need to automate the processing of large volumes of documentation, minimize errors, and increase the efficiency of business processes. The paper analyzes existing solutions and proposes innovative approaches, including the use of machine learning, optical character recognition (OCR), and intelligent search algorithms. The application of AI in document management systems contributes to accelerating information processing, improving data accuracy, and enhancing corporate document management efficiency.

Keywords: electronic document management, artificial i ntelligence, document digitization, machine learning, business process automation

For citation: A. Kassymova, R. Tlegenov, A. Alimagambetova, G. Murzabekova, M. Kassim. Development of document accounting and digitalization systems with integration of artificial intelligence// International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 208–218. (In Kaz.). https://doi.org/10.54309/IJICT.2025.24.4.012

Conflict of interest: The authors declare that there is no conflict of interest.

ҚҰЖАТТАРДЫ ЕСЕПКЕ АЛУ ЖӘНЕ ЦИФРЛАНДЫРУ ЖҮЙЕЛЕРІН ӘЗІРЛЕУ, ЖАСАНДЫ ИНТЕЛЛЕКТТІ ИНТЕГРАЦИЯЛАУМЕН

А.Х. Касымова[1], Р.Н. Тлегенов[1,*], А.З. Алимагамбетова[2], Г.Е. Мырзабекова[3], М. Кассим[4]

[1]Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал, Қазақстан;

[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[3]С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, Астана, Қазақстан;

[4] MARA технологиялар университеті, Малайзия.
E-mail: rinattlegenov1@gmail.com

Акмарал Касымова — қауымдастырылған профессор, педагогика ғылымдарының кандидаты, Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал Қазақстан
E-mail: kasimova_ah@mail.ru,https://orcid.org/0000-0002-4614-4021;
Ринат Тілегенов — магистрант, Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал, Қазақстан
https://orcid.org/0009-0004-5476-5970;
Айнагуль Алимагамбетова — Л.Н.Гумилев атындағы Еуразия ұлттық университеті, «Ақпараттық жүйелер» кафедрасының аға оқытушысы, физика-математика ғылымдарының кандидаты, Астана, Қазақстан
https://orcid.org/0000-0002-9859-2029;
Гүлден Мырзабекова — С.Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті, «Компьютерлік ғылымдар» кафедрасының доценті, физика-

математика ғылымдарының кандидаты, Астана, Қазақстан
https://orcid.org/0000-0001-9807-5200;
Кассим Муризах — PhD, қауымдастырылған профессор, MARA технологиялар университеті, Малайзия
https://orcid.org/0000-0002-8494-4783.

Аннотация. Бұл мақалада жасанды интеллект (ЖИ) технологияларын пайдалану арқылы құжаттарды есепке алу және цифрландыру жүйелерін әзірлеу мен енгізу мәселелері қарастырылады. Тақырыптың өзектілігі құжат айналымының үлкен көлемін автоматтандыру қажеттілігімен, қателерді азайту және бизнес-процестердің тиімділігін арттырумен айқындалады. Жұмыста бар шешімдер талданып, машиналық оқыту, оптикалық таңбаларды тану (OCR) және интеллектуалды іздеу алгоритмдерін қолдануды қамтитын инновациялық тәсілдер ұсынылады. Құжаттарды есепке алу жүйелерінде жасанды интеллектіні қолдану ақпаратпен жұмыс істеу жылдамдығын арттыруға, деректердің дәлдігін қамтамасыз етуге және корпоративтік құжаттаманы басқаруды жақсартуға мүмкіндік береді.

Түйін сөздер: электрондық құжат айналымы, жасанды интеллект, құжаттарды цифрландыру, машиналық оқыту, бизнес-процестерді автоматтандыру

Дәйексөздер үшін: А.Х. Касымова, Р.Н. Тлегенов, А.З. Алимагамбетова, Г.Е. Мырзабекова, М. Кассим. Құжаттарды есепке алу және цифрландыру жүйелерін әзірлеу, жасанды интеллектті интеграциялаумен // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 208–218 бет. (Қазақ тіл). https://doi.org/10.54309/IJICT.2025.24.4.012.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА СИСТЕМ УЧЕТА И ОЦИФРОВКИ ДОКУМЕНТОВ С ИНТЕГРИРОВАНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А.Х.Касымова[1], Р.Н. Тлегенов[1,*], А.З. Алимагамбетова[2], Г.Е. Мырзабекова[3], . Кассим[4]

[1] Западно-Казахстанский аграрно-технический университет им. Жангир хана, Уральск;

[2] Евразийский национальный университет имени Л.Н.Гумилева, Астана, Казахстан;

[3]Казахский агротехнический научно-исследовательский университет имени С. Сейфуллина, Астана, Казахстан;

[4]Технологический университет MAPA, Малайзия.

E-mail: rinattlegenov1@gmail.com

Акмарал Касымова — доцент, кандидат педагогических наук, Западно-Казахстанский аграрно-технический университет имени Жангир хана, г. Уральск, Казахстан
E-mail: kasimova_ah@mail.ru, https://orcid.org/0000-0002-4614-4021;
Ринат Тлегенов — магистрант,западно-Казахстанский аграрно-технический университет имени Жангир хана,г. Уральск, Казахстан
https://orcid.org/0009-0004-5476-5970;
Айнагуль Алимагамбетова — Евразийский национальный университет имени Л.Н.Гумилева, старший преподаватель кафедры «Информационные системы», кандидат физико-математических наук, Астана, Казахстан
https://orcid.org/0000-0002-9859-2029;
Гүлден Мырзабекова — Казахский агротехнический научно-исследовательский университет имени С.Сейфуллина, доцент кафедры компьютерных наук, кандидат физико-математических наук, Астана, Казахстан
https://orcid.org/0000-0001-9807-5200;
Кассим Муризах — Доктор философии, доцент, Технологический университет MARA, Малайзия
https://orcid.org/0000-0002-8494-4783.

Аннотация. В данной статье рассматриваются вопросы разработки и внедрения систем учета и цифровизации документов с использованием технологий искусственного интеллекта (ИИ). Актуальность темы обусловлена необходимостью автоматизации обработки большого объема документации, минимизации ошибок и повышения эффективности бизнес-процессов. В работе анализируются существующие решения, предлагаются инновационные подходы, включая использование машинного обучения, оптического распознавания символов (OCR) и интеллектуальных поисковых алгоритмов. Применение ИИ в системах учета документов способствует ускорению работы с информацией, повышению точности данных и улучшению управления корпоративной документацией.

Ключевые слова: электронный документооборот, искусственный интеллект, оцифровка документов, машинное обучение, автоматизация бизнес-процессов

Для цитирования: А.Х.Касымова, Р.Н. Тлегенов, А.З.Алимагамбетова, Г.Е. Мырзабекова, М. Кассим. Разработка систем учета и оцифровки документов с интегрированием искусственного интеллекта//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 208–218. (На **каз**.). https://doi.org/10.54309/IJICT.2025.24.4.012.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Кіріспе

Қазіргі таңда қоғам мен бизнестің цифрлық трансформациясы құжаттарды есепке алу және цифрландыру жүйелерін деректерді басқарудың тиімді құралына айналдырды (IDC, 2021; Gartner, 2022; ҚР ЦДИАӨМ, 2023). Жасанды интеллект (ЖИ) технологияларын осындай жүйелерге енгізу құжаттармен жұмыс істеу процесін автоматтандыруға, адами фактордың әсерін азайтуға және ақпаратты өңдеу дәлдігін арттыруға мүмкіндік береді (Deloitte, 2022). Бұл зерттеуде жасанды интеллекті пайдалану арқылы құжаттарды есепке алу және цифрландыру жүйелерін әзірлеу мен енгізудің қазіргі заманғы тәсілдері қарастырылады. Қазіргі ұйымдар бухгалтерлік есептер, заңдық актілер, клиенттік келісімшарттар және өзге де құжаттардың үлкен көлемін өңдейді. Мұндай құжаттарды қолмен өңдеу көп уақытты қажет етеді және қателік ықтималдығын арттырады. Сондықтан жасанды интеллект технологияларын құжат айналымында пайдалану ерекше өзектілікке ие. Мұндай технологиялар құжаттарды өңдеу мен талдау жылдамдығын арттырып, рутинді процестерді автоматтандыруға, қателер санын азайтуға және деректердің қауіпсіздігін қамтамасыз етуге мүмкіндік береді. Зерттеу жұмысының мақсаты – жасанды интеллект технологияларын интеграциялау арқылы құжаттарды есепке алу және цифрландырудың интеллектуалды жүйесін құру және оның тиімділігін арттыру жолдарын айқындау. Бұл мақсатқа жету үшін қазіргі қолданыстағы жүйелер мен технологиялар талданып, ЖИ қолдану мүмкіндіктері зерттеледі және жүйеге қойылатын негізгі техникалық және бағдарламалық талаптар айқындалады.

Зерттеу жаңалығы – құжат айналымы мен деректерді басқару саласында жасанды интеллект элементтерін кешенді қолдану арқылы автоматтандырудың жаңа моделін ұсынуында. Ұсынылған тәсіл құжаттарды талдау мен құрылымдау үшін машиналық оқыту алгоритмдерін, қағаз түріндегі материалдарды цифрлық форматқа көшіру үшін оптикалық таңбаларды тану (OCR) технологиясын және мәтін мазмұнын түсіну үшін табиғи тілді өңдеу (NLP) әдістерін қолдануды қамтиды. Сонымен қатар, интеллектуалды іздеу мен чат-боттар сияқты өзара әрекеттесу құралдары пайдаланушыға жүйемен ыңғайлы жұмыс істеуге мүмкіндік береді. Гибридті жасанды интеллект үлгілерін қолдану (нейрондық желілер мен дәстүрлі алгоритмдерді біріктіру) құжаттарды автоматты түрде жіктеудің дәлдігін арттырады, ал семантикалық талдау мен болжамды модельдер деректердің құрылымын тереңірек түсінуге және ұйымдардың ақпараттық процестерін тиімді басқаруға жол ашады. Осылайша, жасанды интеллектке негізделген құжаттарды есепке алу және цифрландыру жүйелерін әзірлеу қазіргі заманғы кәсіпорындардың өнімділігін арттыру, ресурстарды үнемдеу және ақпараттық қауіпсіздікті қамтамасыз ету бағытындағы маңызды қадам болып табылады.

Әдістер мен материалдар

Зерттеу нәтижелері құжаттарды есепке алу және цифрландыру жүй-

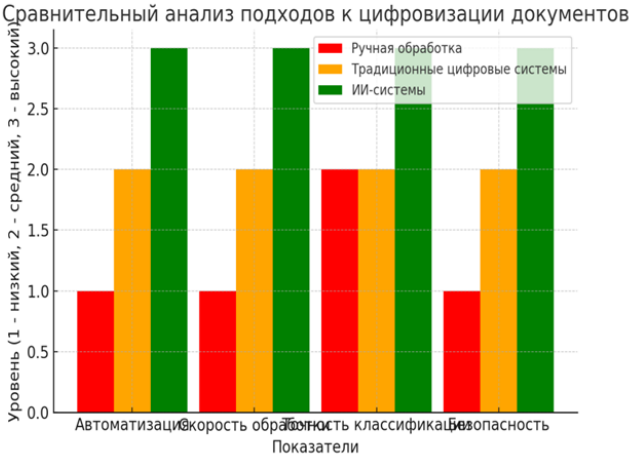
елерінің әртүрлі тәсілдерін салыстырмалы талдау арқылы бағаланды. Әрбір тәсіл автоматтандыру деңгейі, өңдеу жылдамдығы, классификация дәлдігі және деректердің қауіпсіздігі бойынша қарастырылды.

1-кестеде құжаттарды есепке алу мен цифрландыруға арналған тәсілдердің салыстырмалы талдауы келтірілген. Бұл кесте әр әдістің тиімділік деңгейін айқындауға және жасанды интеллект негізіндегі интеллектуалды жүйелердің артықшылықтарын көрсетуге мүмкіндік береді.

Тәсіл	Автоматтандыру	Өңдеу жылдамдығы	Классификация дәлдігі	Қауіпсіздік
Қолмен өңдеу	Төмен	Төмен	Орташа	Төмен
Дәстүрлі цифрлық жүйелер	Орташа	Орташа	Орташа	Орташа
ЖИ негізіндегі интеллектуалды жүйелер	Жоғары	Жоғары	Жоғары	Жоғары

Кесте-1 деректері көрсеткендей, құжаттарды қолмен өңдеу тиімділігі ең төмен тәсіл болып табылады, себебі ол көп уақытты талап етеді және қателіктерге бейім. Дәстүрлі цифрлық жүйелер белгілі бір дәрежеде автоматтандыруды қамтамасыз етеді, алайда әлі де адам араласуын қажет етеді. Ал жасанды интеллектке негізделген интеллектуалды жүйелер ең жоғары автоматтандыру деңгейін қамтамасыз етіп, құжаттарды өңдеу жылдамдығы мен дәлдігін арттырады, сондай-ақ деректердің қауіпсіздігін күшейтеді.

1-суретте құжаттарды цифрландыру тәсілдерінің салыстырмалы нәтижелері диаграмма түрінде көрсетілген.



Сур. 1. Құжаттарды цифрландыру тәсілдерінің салыстырмалы талдауы

Жүргізілген әдістемелік талдау нәтижесінде ұсынылған жүйе құжаттарды өңдеу мен сақтау процестерін толық автоматтандыруға бағытталғанын

көрсетті. Кесте 1 мен 1-суретте берілген салыстырмалы талдау дәстүрлі тәсілдерге қарағанда жасанды интеллект негізіндегі жүйелердің өнімділігі мен тиімділігінің айтарлықтай жоғары екенін дәлелдейді.

Ұсынылған әдістеме келесі кезеңдерді қамтиды:

1. Құжаттардың бастапқы деректерін цифрлық форматқа түрлендіру;
2. OCR және NLP технологиялары арқылы мәтіндік ақпаратты тану және құрылымдау;
3. Машиналық оқыту алгоритмдерін пайдалана отырып, құжаттарды автоматты түрде санаттарға бөлу және талдау;
4. Алынған деректер негізінде интеллектуалды іздеу және болжау жүйелерін іске асыру.

Бұл кезеңдер құжат айналымын оңтайландыруға, уақыт шығынын азайтуға және ақпараттың нақтылығын арттыруға бағытталған. Осылайша, «Әдістер мен материалдар» бөлімінде сипатталған тәсіл зерттеу жұмысының практикалық негізін құрайды және келесі бөлімде ұсынылатын нәтижелерді алуға мүмкіндік берді. Зерттеу барысында OCR, NLP және машиналық оқыту әдістері пайдаланылды (Mithe et al., 2020; Afzal et al., 2021; Xu et al., 2020). Табиғи тілді өңдеу (NLP) әдістері мәтіндік құжаттардың мазмұнын автоматты түрде талдау, негізгі ұғымдарды анықтау және деректерді құрылымдау үшін қолданылды (Devlin et al., 2019; Искаков, 2021).

Нәтижелер және талқылау.

Зерттеу нәтижелері көрсеткендей, жасанды интеллект (ЖИ) технологияларын құжаттарды есепке алу және цифрландыру процесіне енгізу құжат айналымының тиімділігін едәуір арттыруға мүмкіндік береді (Deloitte, 2022). Мұндай тәсіл ұйымдардың ақпараттық ресурстарын жылдам өңдеуге, құжаттардың сақталуын қамтамасыз етуге және адами факторға байланысты қателіктерді азайтуға бағытталған.

ЖИ технологиялары құжаттарды цифрландырудың түрлі аспектілерін автоматтандыруда кеңінен қолданылады. Атап айтқанда:

– Табиғи тілді өңдеу (NLP) мәтінді автоматты түрде талдауға, аннотациялауға және құрылымдауға мүмкіндік береді (Ким және Исмаилова, 2022).;

– Оптикалық таңбаларды тану (OCR) қағаз құжаттарды іздеу және талдау мүмкіндігі бар цифрлық форматқа түрлендіреді (Gonzalez және Woods, 2018).;

– Машиналық оқыту (ML) үлкен деректер жиынтығын талдап, құжаттардың категорияларын, негізгі тақырыптарын және өңдеу басымдықтарын болжайды;

– Кластерлеу және тақырыптық талдау ұқсас құжаттарды автоматты түрде топтастыруды және негізгі тақырыптарды анықтауды қамтамасыз етеді;

– Автоматты деректерді шығару құрылымдалмаған мәтіндерден (мысалы, күндер, атаулар, сома) маңызды ақпаратты бөліп алуға мүмкіндік береді;

– Интеллектуалды іздеу контекстік және семантикалық байланыстарды ескере отырып, қажетті құжаттарды жылдам табуға жағдай жасайды;

– Болжамдық аналитика деректер негізінде тәуекелдерді болжауға және шешім қабылдау процестерін автоматтандыруға мүмкіндік береді.

Мысал ретінде шот-фактураларды автоматты өңдеу жүйесін алуға болады. Мұндай жүйеде машиналық оқыту алгоритмдері құжаттарды автоматты түрде талдап, оларды санаттар бойынша (бухгалтерлік, заңдық, клиенттік) жіктейді, негізгі деректерді (сома, күн, контрагент) шығарып, тиісті бөлімдерге бағыттайды. Нәтижесінде құжаттарды өңдеу уақыты шамамен 70 % қысқарады, ал адами қатеге байланысты тәуекелдер айтарлықтай төмендейді.

Құжаттарды есепке алу мен цифрландырудың интеллектуалды жүйесінің концепциясы.

Ұсынылған жүйе бірнеше өзара байланысты модульдерден тұрады:

1. OCR модулі – қағаз түріндегі құжаттарды цифрлық форматқа түрлендіреді және мәтіндік талдауды қамтамасыз етеді;
2. NLP модулі – мәтін мазмұнын түсініп, автоматты классификация мен маңызды ақпаратты шығару қызметін атқарады;
3. ML модулі – құжаттарды талдау, категориялау, аномалияларды анықтау және автоматты бағыттау үшін қолданылады;
4. Интеллектуалды іздеу жүйесі – семантикалық талдау негізінде ақпаратты жылдам табуға мүмкіндік береді;
5. Қауіпсіздік және қолжетімділік жүйесі – пайдаланушы құқықтарын бақылап, деректерді шифрлау және қорғау функцияларын орындайды;
6. Бизнес-процестерді автоматтандыру модулі – корпоративтік жүйелермен (ERP, CRM) интеграцияланып, нұсқаларды басқаруды және құжат айналымын автоматтандырады (Кожамкулова және Козбагаров, 2022);
7. Адаптивті өзін-өзі оқыту алгоритмдері – уақыт өте келе жүйенің дәлдігін арттырады;
8. Блокчейн модулі – құжаттардың тұтастығын және өзгерістер тарихының сенімділігін қамтамасыз етеді.

Жүйенің техникалық және бағдарламалық талаптары

Техникалық талаптар:

– Жоғары өнімді серверлік инфрақұрылым және бұлттық сақтау жүйелері (Markets and Markets, 2022);

– Қолданыстағы корпоративтік жүйелермен (ERP, CRM, ECM) интеграция мүмкіндігі (Nurzhanov және Tulegenov, 2021);

– API арқылы сыртқы сервистермен өзара әрекеттесу қолдауы;

– Деректердің қауіпсіздігі мен сақтық көшірмесін қамтамасыз ету;

– Машиналық оқытуға арналған GPU/TPU аппараттық жеделдеткіштерін пайдалану;

– Жүйені деректер көлеміне қарай масштабтау мүмкіндігі.

Бағдарламалық талаптар:

– Машиналық оқыту алгоритмдерін (нейрондық желілер, классификаторлар, бустинг) қолдау;

– NLP және OCR технологияларын пайдалану;

– Қол жеткізу мен қауіпсіздік жүйесін іске асыру (аутентификация, шифрлау, аудит);

– Блокчейнмен интеграция арқылы құжаттардың өзгермейтіндігін қамтамасыз ету (Садыков және Нурлыбаев, 2023);

– Пайдаланушы интерфейсін бейімдеу және көппайдаланушылық режимді қолдау;

– Семантикалық талдау негізінде интеллектуалды іздеу жүйесін енгізу.

Кесте.2. OCR, NLP, машиналық оқыту және блокчейн технологияларының салыстырмалы сипаттамасы.

Технология	Негізгі мақсаты	Артықшылықтары	Шектеулері
OCR (оптикалық таңбаларды тану)	Қағаз түріндегі құжаттарды цифрлық форматқа түрлендіру	Мәліметтерді енгізуді автоматтандыру, қол еңбегін азайту, өңдеу жылдамдығының жоғары болуы	Суреттер сапасы төмен болған жағдайда дәлдіктің төмендеуі
NLP (табиғи тілді өңдеу)	Мәтіндік ақпаратты талдау және түсіну	Құжаттарды автоматты түрде классификациялау, интеллектуалды іздеу, контекстті талдау	Көпмағыналы мәтіндерді өңдеудің күрделілігі
Машиналық оқыту (ML)	Мәліметтерді автоматты түрде өңдеу және талдау	Заңдылықтарды анықтау, болжау, өзін-өзі үйрету қабілеті	Үлкен көлемдегі деректерді оқыту үшін қажет етуі
Блокчейн	Мәліметтердің қауіпсіздігі мен өзгеріссіздігін қамтамасыз ету	Қорғаудың жоғары деңгейі, ашықтық, децентрализация	Есептеу шығындарының жоғары болуы, интеграцияның күрделілігі

Жүргізілген талдау нәтижелері көрсеткендей, ЖИ технологияларын біріктіре отырып құжат айналымын автоматтандыру дәстүрлі жүйелерге қарағанда тиімдірек. Ұсынылған жүйе деректерді өңдеу жылдамдығын арттырады, құжаттардың дұрыстығын қамтамасыз етеді және ұйымдық процестердің ашықтығын күшейтеді. Сонымен қатар, блокчейн мен семантикалық іздеу әдістерін қолдану құжаттық ақпараттың тұтастығын қамтамасыз етіп, деректердің бұрмалануына жол бермейді.Осылайша, зерттеу нәтижелері ұсынылған концепцияның практикалық маңыздылығын дәлелдейді және Қазақстандағы құжат айналымының цифрлық трансформациясын дамытуға үлес қосады.

Қорытынды

Бұл зерттеу Sentinel-2 спутниктік деректері мен ERA5-Land климаттық параметрлерін пайдалана отырып, топырақтың тұздылығын болжау және сегменттеу мәселесіне арналған кешенді тәсілдің тиімділігін көрсетті. Әзірленген модель бақыланбайтын әдістерді (KMeans, агломеративті кластерлеу, DBSCAN), бақыланатын оқытуды (XGBoost) және кеңістіктік гетерогенді деректермен жұмыс істеу кезінде жоғары дәлдік пен сенімділікке

қол жеткізген көп тапсырмалы нейрондық желіні біріктірді.

Негізгі кластерлеу сапа көрсеткіштері әдістің тиімділігін айқын дәлелдеді: Silhouette ұпайы 0,8156-ға жетті, ал Davies–Bouldin ұпайы 0,3022-ге дейін төмендеді. Классификациялық тапсырмаларда модельдің дәлдігі 99,99% құрап, 10 000-нан астам сынақ үлгілерінің ішінде тек бір ғана қате тіркелді.

Ұсынылған әдіс топырақтың әртүрлі типтерінің спектрлік ерекшеліктерін бейімдеуге, шекаралық қателіктерді азайтуға және сортаңдану процестеріне климаттық факторлардың әсерін ескеруге мүмкіндік береді. Ансамбльдік тәсіл мен ықтималдықтық белгілерді қолдану дәстүрлі алгоритмдермен салыстырғанда топырақ кластарын неғұрлым дәл бөлуді қамтамасыз етті.

Алынған нәтижелер ұсынылған тәсілдің ғылыми және қолданбалы маңыздылығын растайды. Бұл әдіс жердің тозуын бақылау, топырақ құнарлылығының төмендеу қаупін бағалау және агроэкологиялық шешімдерді қолдау мақсатында пайдалануға болады.

Болашақта әдістемені жетілдіру бағыттары ретінде конволюциялық нейрондық желілерді қолдану арқылы терең сегменттеу тәсілін дамыту, ұшқышсыз әуе аппараттарының деректерімен интеграциялау және талдау ауқымын басқа аймақтар мен маусымдық кезеңдерге кеңейту жоспарлануда. Мұндай жетілдірулер климаттың өзгеруі жағдайында топырақтың тұзденуын мониторингтеудің әмбебап және масштабталатын жүйесін қалыптастыруға мүмкіндік береді.

REFERENCES

Afzal M.Z., Kölsch A., Ahmed S., Liwicki M. (2021). Document Classification with Deep Learning: Recent Advances and Challenges // International Conference on Document Analysis and Recognition (ICDAR). — 2021. — Pp. 478–486.

Deloitte. (2022). Automation with intelligence: Pursuing organisation-wide reimagination. — 2022. — URL: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Technology/gx-automation-with-intelligence.pdf

Devlin J., Chang M.W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. — 2019. — Pp. 4171–4186.

Gartner. (2022). Market Guide for Intelligent Document Processing Solutions. — 2022. — URL: https://www.gartner.com/en/documents/4010955

Gonzalez R.C., Woods R.E. (2018). Digital Image Processing. — 4th ed. — Pearson, 2018.

IDC. (2021). Worldwide Global Data Sphere Forecast, 2021–2025: The World Keeps Creating More Data. — 2021. — URL: https://www.idc.com/getdoc.jsp?containerId=US46410421

Искаков К.С. (2021). Заң құжаттарын талдау үшін табиғи тілді өңдеу әдістері // ҚазҰУ Хабаршысы. Ақпараттық технологиялар сериясы. — 2021. — №3. — 102–110 бб.

Ким Д.К., Исмаилова А.А. (2022). Қазақстандағы электрондық құжат айналымы жүйелерінде екі тілді құжаттарды танудың ерекшеліктері // ҚазҰУ Хабаршысы. Ақпараттық технологиялар сериясы. — 2022. — №2. — 78–85 бб.

Қазақстан Республикасының Цифрлық даму, инновациялар және аэроғарыш өнеркәсібі министрлігі. (2023). Қазақстан Республикасындағы цифрландырудың жай-күйі туралы ұлттық баяндама. — 2023.

Қожамқұлова Ж.Т., Қозбағаров О.А. (2022). Электрондық құжат айналымы жүйелеріндегі деректерді сақтаудың гибридті үлгілері // ҚазҰТЗУ Хабаршысы. — 2022. — №4. — 156–162 бб.

Markets and Markets. (2022). Cloud Content Management Market – Global Forecast to 2027. — 2022. — URL: https://www.marketsandmarkets.com/Market-Reports/cloud-content-management-market-83195810.html

Mithe R., Indalkar S., Divekar N. (2020). Optical Character Recognition // International Journal of Recent Technology and Engineering. — 2020. — Vol. 9. — No. 1. — Pp. 2277–3878.

Nurzhanov B., Tulegenov E. (2021). Cloud-based Document Management Systems in Kazakhstan: Challenges and Perspectives // International Journal of Computer Applications. — 2021. — Vol. 183. — No. 8. — Pp. 32–38.

Садыков Т.Р., Нұрлыбаев К.А. (2023). Электрондық құжат айналымы жүйелерін мемлекеттік ақпараттық жүйелермен біріктіру // Қазақстандағы ақпараттық технологиялар. — 2023. — №1. — 45–51 бб.

Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding // Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. — 2020. — Pp. 1192–1200.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 219–238
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.013

УДК: 004.85, 004.56.

ANALYSIS OF RECOGNITION ALGORITHMS AND CONVOLUTIONAL NEURAL NETWORK FOR HAND GESTURE RECOGNITION IN KAZAKH SIGN LANGUAGE

N.N.Les[1*], S.B.Mukhanov[2], M.T.Ipalakova[1], A.K.Mustafina[1]
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Astana IT University, Astana, Kazakhstan.
E-mail: nurzhaan0@gmail.com

Nurzhan Les — Master of technical science, International Information Technology University, Almaty, Kazakhstan
E-mail: nurzhaan0@gmail.com, https://orcid.org/0009-0008-2909-3606;
Samat Mukhanov — PhD in Computer systems and software engineering, Assistant-professor of Cybersecurity School, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-8761-4272;
Madina Ipalakova — Candidate of technical sciences, Associate Professor, Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-8700-1852;
Akkyz Mustafina — Candidate of technical sciences, Vice-rector for academic affairs, International Information Technology University, Almaty, Kazakhstan
https://orcid.org/0000-0002-5884-9280.

Abstract. The ever-growing interest in machine learning and neural networks is fueled by significant advancements in computational capabilities, enabling breakthroughs in object, sound, text, and other forms of data recognition. These advancements have paved the way for a more intuitive interaction between humans and machines, making such technologies accessible to a wider audience. Recent developments in computer vision, in particular, have led to the creation of sophisticated models capable of recognizing objects in images and videos. This same technology has been effectively adapted for hand gesture recognition, enabling applications in fields like human-computer interaction, robotics, and sign language interpretation. This paper explores some of the most popular hand gesture recognition models, with a particular focus on Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs). These models differ in their methodologies, processing efficiency, and the volume of training data they require, offering various advantages and limitations depending on the application context. The core objective of this study is to provide an overview of diverse machine learning algorithms, delving deeply into their

theoretical underpinnings, operational mechanisms, and comparative performance in terms of accuracy, training time, and data requirements. In addition, this work presents experimental results of sign language recognition in Kazakh, specifically using the dactyl alphabet. A detailed analysis is provided, accompanied by a comprehensive table that reports the accuracy of each recognized gesture. Real-time testing was conducted with individual hand gestures displayed in front of a camera, showcasing the effectiveness of the recognition system. Furthermore, the study incorporates an explanation of the mathematical foundations and logical structures underlying machine learning algorithms, illustrated through formulae, functional relationships, and flowcharts that depict the recognition process. By combining theoretical insights with practical experiments, this paper aims to contribute to the growing field of gesture recognition and its applications in accessible communication technologies.

Keywords: Hand gesture recognition, neural networks, algorithm, layer, CNN, SVM, YOLO

For citation: N.N. Les, S.B. Mukhanov, M.T. Ipalakova, A.K. Mustafina. Analysis of recognition algorithms and convolutional neural network for hand gesture recognition in kazakh sign language//International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 219–238. (In Eng.). https://doi. org/10.54309/IJICT.2025.24.4.013.

Conflict of interest: The authors declare that there is no conflict of interest.

ҚАЗАҚ ТІЛІНДЕ ҚОЛ ҚИМЫЛДАРЫН ТАНУ ҮШІН ТАНУ АЛГОРИТМДЕРІН ЖӘНЕ КОНВОЛЮЦИЯЛЫҚ НЕЙРОНДЫҚ ЖЕЛІНІ ТАЛДАУ

Н.Н. Лес[1*], С.Б. Муханов[2], М.Т. Ипалакова[1], А.К. Мустафина[1]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2]Астана IT университеті, Астана, Қазақстан;
E-mail: nurzhaan0@gmail.com

Нуржан Лес — Магистрант, «Бағдарламалық қамтамасыз ету Инженериясы» білім беру бағдарламасы, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан
E-mail: nurzhaan0@gmail.com, https://orcid.org/0009-0008-2909-3606;
Самат Муханов — PhD, Киберқауіпсіздік мектебінің ассистент-профессоры, Астана IT университеті, Астана, Қазақстан
https://orcid.org/0000-0001-8761-4272;
Мадина Ипалакова —Техника ғылымдарының кандидаты, Халықаралық ақпараттық технологиялар университетінің Компьютерлік инженерия кафедрасының профессоры, Алматы, Қазақстан
https://orcid.org/0000-0002-8700-1852;
Аккыз Мустафина — Техника ғылымдарының кандидаты, Халықаралық ақпараттық технологиялар университетінің академикалық сұрақтар жөніндегі Проректоры, Алматы, Қазақстан
https://orcid.org/0000-0002-5884-9280.

Аннотация. Машиналық оқыту мен нейрондық желілерге деген қызығушылықтың үнемі артуы есептеу мүмкіндіктерінің айтарлықтай жетістіктерімен қамтамасыз етіледі, бұл объектілік, дыбыстық, мәтіндік және деректерді танудың басқа түрлерінде жетістіктерге жетуге мүмкіндік береді. Бұл жетістіктер адамдар мен машиналар арасындағы интуитивті өзара әрекеттесуге жол ашып, мұндай технологияларды кең аудиторияға қолжетімді етті. Компьютерлік көрудің соңғы жетістіктері, атап айтқанда, кескіндер мен бейнелердегі объектілерді тануға қабілетті күрделі модельдердің жасалуына әкелді. Дәл осы технология адам мен компьютердің өзара әрекеттесуі, робототехника және жестау тілін түсіндіру сияқты салаларда қолдануға мүмкіндік беретін қол қимылдарын тануға тиімді бейімделген. Бұл мақалада Конволюциялық Нейрондық Желілерге (Cnn) және Тірек Векторлық Машиналарға (Svm) ерекше назар аудара отырып, қол қимылдарын танудың ең танымал үлгілері қарастырылады. Бұл модельдер әдістемелерімен, өңдеу тиімділігімен және қажетті оқу деректерінің көлемімен ерекшеленеді, қолдану контекстіне байланысты әртүрлі артықшылықтар мен шектеулерді ұсынады. Бұл зерттеудің негізгі мақсаты-машиналық оқытудың әртүрлі алгоритмдеріне шолу жасау, олардың теориялық негіздерін, операциялық механизмдерін және салыстырмалы өнімділігін дәлдік, оқу уақыты және деректерге қойылатын талаптар тұрғысынан терең зерттеу. Сонымен қатар, бұл жұмыста дактил алфавитін қолдана отырып, signay тілін қазақ тілінде танудың эксперименттік нәтижелері келтірілген. Егжей-тегжейлі талдау әрбір танылған қимылдың дәлдігі туралы есеп беретін жан-жақты кестемен бірге беріледі. Нақты уақыттағы тестілеу тану жүйесінің тиімділігін көрсететін камера алдында көрсетілген жеке қол қимылдарымен жүргізілді. Сонымен қатар, зерттеу тану процесін бейнелейтін формулалар, функционалдық қатынастар және блок-схемалар арқылы суреттелген машиналық оқыту алгоритмдерінің негізінде жатқан математикалық негіздер мен логикалық құрылымдарды түсіндіруді қамтиды. Теориялық түсініктерді практикалық эксперименттермен үйлестіре отырып, бұл жұмыс қимылдарды танудың өсіп келе жатқан саласына және оны қолжетімді коммуникациялық технологияларда қолдануға үлес қосуға бағытталған.

Түйін сөздер: Қол қимылдарын тану, нейрондық желілер, алгоритм, қабат, CNN, SVM, YOLOД

әйексөз үшін: Н.Н. Лес, С.Б. Муханов, М.Т. Ипалакова, А.К. Мустафина. Казақ тілінде қол қимылдарын тану үшін тану алгоритмдерін және конволюциялық нейрондық желіні талдау//Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 219–238 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.013.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

АНАЛИЗ АЛГОРИТМОВ РАСПОЗНАВАНИЯ И СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ РАСПОЗНАВАНИЯ ЖЕСТОВ РУК НА КАЗАХСКОМ ЯЗЫКЕ ЖЕСТОВ

Н.Н. Лес[1*], С.Б. Муханов[2], М.Т. Ипалакова[1], А.К. Мустафина[1]

[1]Международный университет информационных технологий, Алматы,

Казахстан;
[2]Астана IT университет, Астана, Казахстан.
E-mail: nurzhaan0@gmail.com

Нуржан Лес — Магистрант, образовательная программа «Программная инженерия», Международный университет информационных технологий, Алматы, Казахстан
E-mail: nurzhaan0@gmail.com, https://orcid.org/0009-0008-2909-3606;
Самат Муханов — PhD, Ассистент-профессор школы кибербезопасность, Астана IT Университет, Астана, Казахстан
https://orcid.org/0000-0001-8761-4272;
Мадина Ипалакова — Кандидат технических наук, доцент кафедры вычислительной техники Международного университета информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-8700-1852;
Аккыз Мустафина — Кандидат технических наук, доцент кафедры вычислительной техники Международного университета информационных технологий, Алматы, Казахстан
https://orcid.org/0000-0002-5884-9280.

Аннотация. Постоянно растущий интерес к машинному обучению и нейронным сетям подогревается значительными достижениями в области вычислительных возможностей, позволяющими совершать прорывы в распознавании объектов, звука, текста и других форм данных. Эти достижения проложили путь к более интуитивному взаимодействию между людьми и машинами, сделав такие технологии доступными для более широкой аудитории. Последние разработки в области компьютерного зрения, в частности, привели к созданию сложных моделей, способных распознавать объекты на изображениях и видео. Эта же технология была эффективно адаптирована для распознавания жестов рук, что позволяет применять ее в таких областях, как взаимодействие человека и компьютера, робототехника и сурдоперевод. В этой статье рассматриваются некоторые из наиболее популярных моделей распознавания жестов рук, с особым акцентом на сверточные нейронные сети (CNN) и методы опорных векторов (SVM). Эти модели различаются по своим методологиям, эффективности обработки и объему требуемых обучающих данных, предлагая различные преимущества и ограничения в зависимости от контекста применения. Основная цель этого исследования - дать обзор различных алгоритмов машинного обучения, глубоко изучив их теоретические основы, операционные механизмы и сравнительную производительность с точки зрения точности, времени обучения и требований к данным. Кроме того, в этой работе представлены экспериментальные результаты распознавания языка жестов на казахском языке, в частности, с использованием дактильного алфавита. Приводится подробный анализ, сопровождаемый подробной таблицей, в которой сообщается о точности каждого распознанного жеста. Тестирование проводилось в режиме реального времени с использованием отдельных жестов рук, отображаемых перед ка-

мерой, что продемонстрировало эффективность системы распознавания. Кроме того, исследование включает в себя объяснение математических основ и логических структур, лежащих в основе алгоритмов машинного обучения, проиллюстрированных формулами, функциональными взаимосвязями и блок-схемами, которые описывают процесс распознавания. Сочетая теоретические идеи с практическими экспериментами, эта статья призвана внести вклад в растущую область распознавания жестов и их применения в доступных коммуникационных технологиях.

Ключевые слова: Распознавание жестов руками, нейронные сети, алгоритм, layer, CNN, SVM, YOLO

Для цитирования: Н.Н. Лес, С.Б. Муханов, М.Т. Ипалакова, А.К. Мустафина. Анализ алгоритмов распознавания и сверточной нейронной сети для распознавания жестов рук на казахском языке жестов//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 219–238. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.013.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

The Recognition of hand gestures (Gesture-recognition) is an important problem in the field of human-computer interaction. Hand gestures are used as a natural and intuitive way of communication between people and digital devices or systems. One of the most significant applications of this field is the recognition of sign language, which is used by people with hearing disabilities for communication. There are many different sign languages used around the world, and one of them is Kazakh Sign Language (KSL). KSL, like many other sign languages, is not widely studied, and there is a need for effective recognition systems to support its users. In recent years, recognition algorithms and deep learning methods, especially Convolutional Neural Networks (CNNs), have achieved significant success in computer vision and pattern recognition tasks. CNNs have shown state-of-the-art performance in various applications, including image classification, object detection, and hand gesture recognition, due to their ability to learn spatial features and hierarchical representations from data. This research paper aims to review and analyze different recognition algorithms and investigate the performance of CNN-based models for hand gesture recognition in Kazakh Sign Language. The goal is to use machine learning and computer vision techniques to improve the accuracy and efficiency of KSL recognition and contribute to the development of assistive technologies for people with hearing disabilities in Kazakhstan.

Review of related literature and problem statement

The state of the art in sign language recognition research is leaning towards deep learning-based systems. The recent works (Wang et al., 2020; Bilgin & Mutludogan, 2019) demonstrate CNN-based frameworks that can reach more than 90 % recognition accuracy on ASL and Turkish Sign Language datasets. However, research efforts dedicated to the problem of Kazakh Sign Language (KSL) recognition are still in their infancy. Kenshimov et al. (2021) and Mukhanov et al. (2023) have paved the way for this topic but used static images for the recognition task. This work extended the pipeline and demonstrated the real-time detection of hand gestures using YOLO

and CNN.

The objective of this work will be to use various algorithms and machine learning methods within the convolutional neural network in order to identify certain gestures of the Kazakh sign language. Therefore, all these algorithms will be considered and tested, and their detailed description will be given. The convolutional neural network is typically used within deep learning for the image recognition set. The training of this network is based on machine learning. In addition to being an interconnected neural network, it is also what it is made of and what metrics it applies for predictions in pattern recognition -in our case, recognizing gestures in the Kazakh sign language. The current work will show the functions used in this convolutional network and the mathematical formulas for them. Convolutional neural networks are a type of deep neural network, which are specifically designed for processing images and videos. CNNs account for the spatial structure between pixels, and convolutional operations are applied to higher-order feature extraction from the data, from simple edges to more complex objects. CNNs have exhibited outstanding performance across many computer vision applications and have also demonstrated potential in other fields. Yann LeCun introduced CNNs in the 1980s, although they did not gain widespread use until the mid-2010s with the emergence of deep learning approaches and the availability of big datasets. CNNs are now widely employed in industry and academia and are anticipated to play an essential role in future technologies.

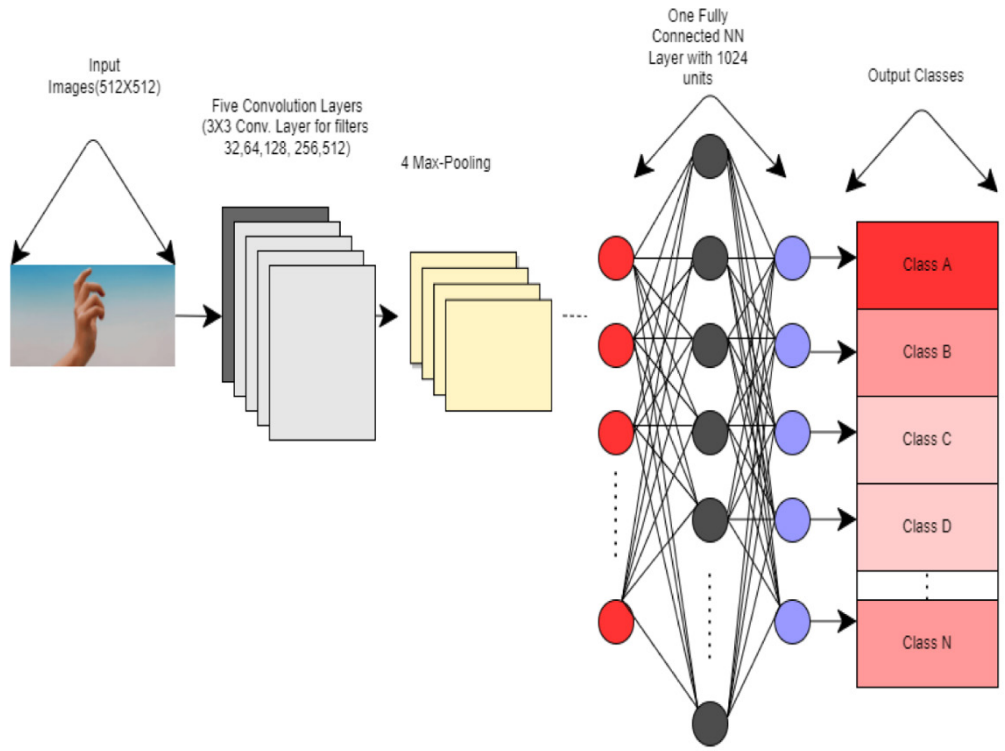


Fig. 1. An example of 2-D convolution

The mathematical operation of two functions (Convolutional operation):

$$[f * g](t) \equiv \int_0^t f(\tau)g(t-\tau)d\tau \quad , \tag{1}$$

where:

$[f * g](t) \equiv$ denotes the convolution of f and g;

The integral runs from 0 to t, integrating over the variable τ;

$f(\tau)$ and $g(t-\tau)$ are the function values at shifted arguments.

A layer will produce as many output feature maps as there are kernels in that layer. The relationship between the size of the output feature map and the input is that each kernel produces one feature map, which describes a single aspect of the input data. The complexity of the task and the data are what will determine the size and number of output feature maps. Each of these feature maps is a two-dimensional array for one filter that describes the presence of that filter in the input data. The values within these filters will be learned through backpropagation in order to get the greatest performance possible for the task at hand.

Alongside these kernels are channels that capture different aspects of the input data. The channels are two-dimensional arrays that each express a single characteristic of the input data. For instance, if the task were classifying images, then the input data might have distinct channels for edge data, textural data, color, and so on. It is the grouping of these channels and kernels that allow for the three key concepts in convolutional neural networks design. These include sparse interactions, sharing of parameters, and equivariant representations. Sparse interactions refers to how each output feature map is the result of convolution of a small kernel with only a few of the input feature maps, as opposed to all of the input feature maps. This also improves performance while also reducing the amount of parameters needed to represent the layer.

This is because one feature map produced by a kernel will have a similar effect; it is therefore sometimes known as parameter sharing. This works in such a way that each output feature map produced in a convolution layer is obtained by using the same kernel. This is akin to saying the weights in that kernel are shared between each of the output feature maps-the model does not learn the weights for each one separately. The immediate consequence of all this is a substantial reduction in the number of parameters needed to represent a model with significant improvements in its generalization performance.

Materials and methods

The CNN method in recognition problems

In this work, YOLO was only responsible for hand detection/localization while another CNN model oversaw classifying the detected gesture. For this, the segmented hand was used as an input for a gesture classifier that was trained separately. This pipeline was verified to be more flexible and allowed for independent fine-tuning of the detector and the classifier.

In this project, we used Convolutional Neural Networks to train machine learning algorithms in the recognition of hand gestures and Kazakh sign language. The results of recognition tested are in this work. However, it is necessary to introduce this network and how it works in pattern recognition. Convolutional Neural Networks (CNNs, in short) are a type of artificial neural network that is highly efficient

at processing and analyzing visual data like images and videos. The basic structure of CNNs consists of an input data, for example an image, to layers of interconnected artificial neurons. Each neuron in each layer of the CNN is only connected to a small patch of the input; these patches are said to overlap, such that the totality of the input data is covered.

The convolution layer, which is one of the most important components of CNNs, makes use of learnable filters called kernels to extract features from the input data. The kernel is moved around the input in a process called convolution. It is worth noting that the kernel multiplies the values that correspond to the input at each location to get a set of feature maps. The number of output feature maps is determined by the number of kernels; this is because every kernel has one output feature map, which describes one particular feature of the input data. It is important to remember that this is the interaction of both these components: channels and kernels that enables sparse interactions, parameter sharing, and equivariant representations-the three important ideas in CNN design.

The choice of YOLO and CNN as recognition architectures was inspired by their mutual strengths. YOLO-based hand detector showed high accuracy and real-time performance while CNN is a state of the art in image classification tasks. Their integration into a single pipeline was found optimal for the task of KSL recognition since hand segmentation and dynamic motion localization are of great importance.

Support Vector Machine for hand gesture recognition

The system for the recognition of hand gestures that show letters is proposed in this paper. This system recognizes the hand gestures using biorthogonal wavelet transformation. The system for recognizing hand gestures is operated in several ways. Firstly, images are read and filtered to clean them from noise. After that, the system recognizes the edges of the images and calculate projections along certain directions by using the Radon transform. After that, the system applies biorthogonal wavelet transformation to these projections. For the recognition of hand gestures the system trains and tests SVM. Support Vector Machine (SVM) is a supervised learning algorithm that can be used for both classification and regression. The basic idea behind an SVM is to find a hyperplane that separates the data in the best possible way in a high-dimensional space. “Support vector” is the name for data points that are nearest to the decision boundary or hyperplane, which are essential to find out the optimal separation. This article present a comparative study of two SVM classificators: “Binary classification” and “Multiclass classification” for hand gesture recognition. In the paper we should like to say a few words about the research of the quality of binary classification. The most attractive properties of binary classification using SVM in hand gesture recognition is its power and efficiency. It has many benefits such as being a robust algorithm, achieving high accuracy, and having an interpretable model. Feature selection and hyperparameter tuning are critical for achieving optimal performance. On the other hand, as the research continues, the more resources of the computer there will be. In terms of application areas, SVMs can be used in various domains such as computer vision, image processing, robotics, and human-computer interaction. The application of binary classification using SVM in real-world scenarios will depend on the specific requirements and constraints of the application. It is also important to note that binary classification using SVM is just one of many algorithms and techniques that can be used for hand gesture recognition, and other meth-

ods may be more suitable depending on the application. We also should like to point out that a second classificator used in the experiment "Multiclass classification". This classificator is not as effective as "Binary classification" and was used in experiment as an example.

Canny Edge Detection Algorithm is the method that system uses for the recognition of the image edges. This algorithm was chosen because this algorithm is the best among many benchmark algorithms. This algorithm finds the most appropriate edges by decreasing error rates, making possible the exact localization of edges, and only marking edges once in case of one edge in order to have a minimal response. The only filter that satisfies the requirements is the first derivative of Gaussian function and that filter can be approximated as.

$$w^T x + b = 0, \tag{2}$$

$$min_{w,b} \frac{1}{2} ||w|| \tag{3}$$

$$y_i(w^T x_i + b) \geq 1, \forall_i, \tag{4}$$

where:

w is the weight vector that defines the hyperplane's orientation, x is the feature vector for the input data, b is the bias term that shifts the decision boundary, y_i is the label of the ith data point (either +1 or -1), and x_i is the feature vector for the ith data point.

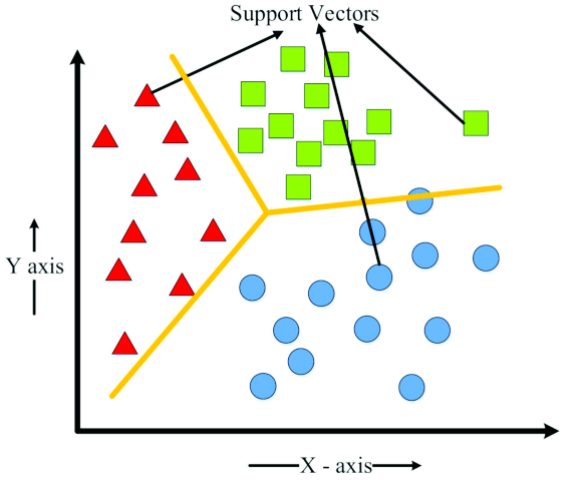


Fig. 2. Binary classification

Multiclass SVMs offer a powerful and versatile approach for hand gesture recognition with multiple classes. Their ability to handle diverse gesture sets and leverage the strengths of binary SVMs makes them valuable for various applications. However, computational complexity and choosing the appropriate multiclass strategy remain key considerations. As research advances and computational resources improve, multiclass SVMs are expected to play a crucial role in expanding the capabilities of hand gesture recognition technologies. Multiclass Support Vector Machines (SVMs) extend the binary classification capabilities of SVMs to recognize multiple hand gestures simultaneously. This is achieved by constructing multiple hyperplanes

in high-dimensional space, each separating a specific class from all other classes.

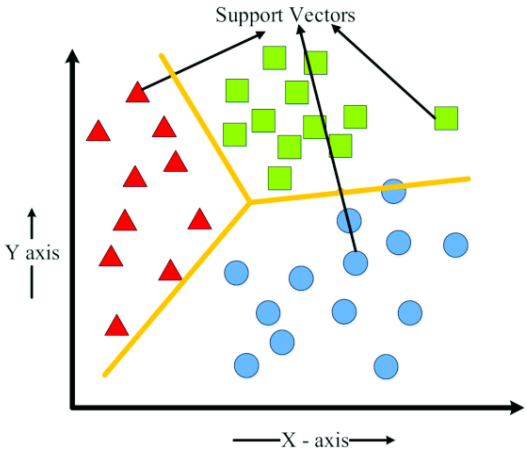


Fig. 3. Multiclass classification

The goal of the project was to create a reliable hand gesture recognition system using support vector machines (SVMs). At the initial stage, a dataset was collected that included 9 different classes of Kazakh sign language. Each of them contains from 150-200 images. Subsequently, segmentation was performed in which the hand gesture region was isolated from the background using a thresholding algorithm. The resulting images were then converted to grayscale, a move aimed at simplifying SVM processing while increasing contrast. Resizing the images to standardized sizes was found to be crucial for more efficient comparisons during training of the SVM model.

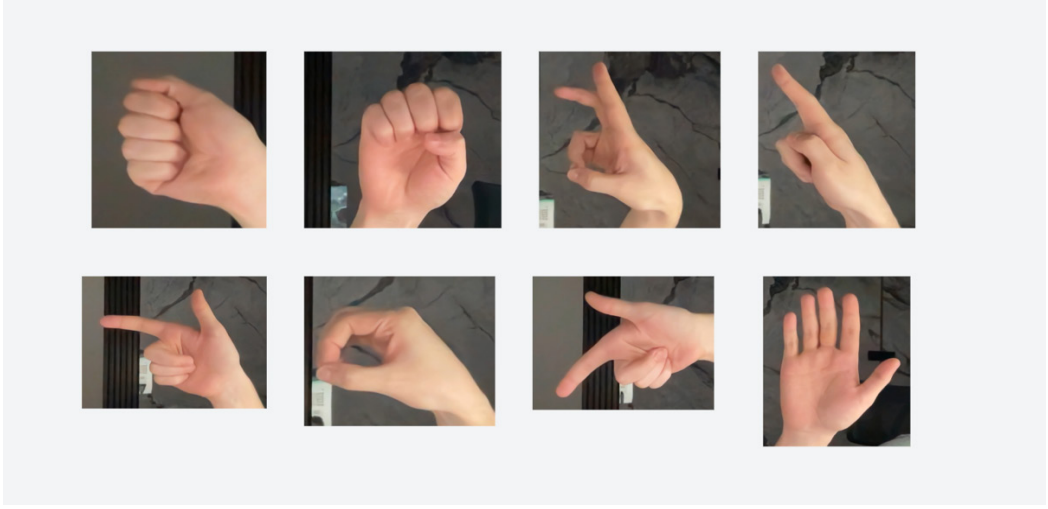


Fig. 4. Finger spelled Alphabet (Top row [A, Aa(Ә), B, D], Bottom row [E, G, Gg(F), V]).

Model training process analysis

Binary classification starts with a relatively low training accuracy of 0.625. However, it shows consistent improvement over the epochs, indicating that the model

learns from the training data and adjusts its predictions accordingly.

The validation accuracy for Binary classification begins at 0.625 and also shows a positive trend over the epochs. However, the fluctuations in the validation accuracy suggest challenges in generalizing the model's predictions to unseen data.

Mutliclass classification starts with a higher training accuracy of 0.635 and steadily increases over the epochs. This indicates that the model effectively learns from the training data and improves its accuracy.

The validation accuracy for multiclass classification begins at 0.635 and follows a positive trend. The close alignment between the training and validation accuracy suggests that Mutliclass classification generalizes well to unseen data.

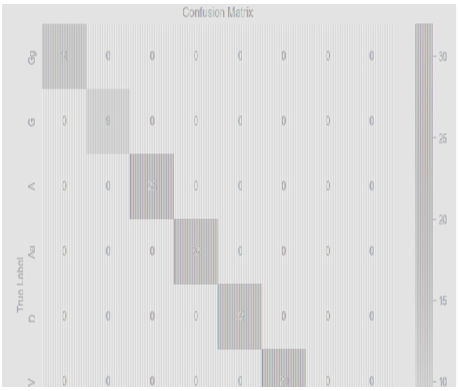


*Fig. 5.*Model accuracy and model loss

Mutliclass classification exhibits higher initial training accuracy and validation accuracy compared to Binary classification. Both models show an improving trend in accuracy over the epochs, indicating effective learning from the training data. However, Mutliclass classification demonstrates better generalization capabilities, as evidenced by its higher validation accuracy and closer alignment between training and validation accuracy.

Analysis of the confusion matrix showed that the model most commonly confused the signs "Г" and "F". The reason is the great similarity in their finger configurations. The second reason for ambiguity is the changing lighting conditions, which adversely affect edge detection in the preprocessing step. Dataset balancing and adaptive preprocessing will be considered in future work to reduce such confusions.

The model still presents high precision, which is indicative of improved reliability in positive predictions.

Recall values are variable across classes, with some classes achieving high recall, such as Blank and Nn(Ң), while others may benefit from further optimization, such as Hh(һ).

F1 scores are a balanced measure of precision and recall, reflecting overall improvement.

The macro average gives the general view of model performance, whereas the weighted average takes into account the class imbalance.

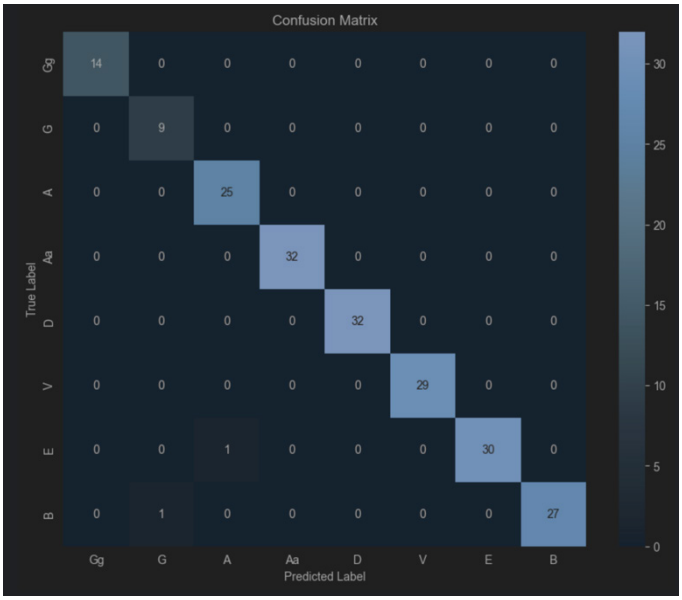


Fig. 6. Confusion matrix

Convolutional neural network with Yolo model

The YOLO (You Only Look Once) model, built on the foundation of convolutional neural networks (CNNs), revolutionized object detection by emphasizing speed, accuracy, and computational efficiency. Unlike traditional methods such as R-CNN and Fast R-CNN, which rely on region proposal networks to first generate potential object locations and then classify each region individually, YOLO operates as a single, end-to-end neural network. It divides the input image into an S×S grid, where each grid cell predicts bounding boxes, object confidence scores, and class probabilities simultaneously.

Key Formulas in YOLO

1) Grid-based Predictions:

The image is divided into a grid of $S \times S$. Each grid cell predicts:

- B: Bounding boxes, each defined by 4 coordinates: (x, y, w, h).
- C: Class probabilities for the
- N object categories.
- P obj: Objectness score, which represents the confidence that an object exists in the grid cell.

The output tensor for YOLO is therefore:

$$S \times S \times (B * 5 + C), \tag{5}$$

Here, 5 accounts for the four bounding box parameters (x, y, w, h) and the confidence score P_{objc}

2) Bounding Box Prediction:

YOLO predicts bounding boxes as offsets relative to the grid cell. For a grid

cell located at (i,j), the bounding box center coordinates (x,y) are given by:

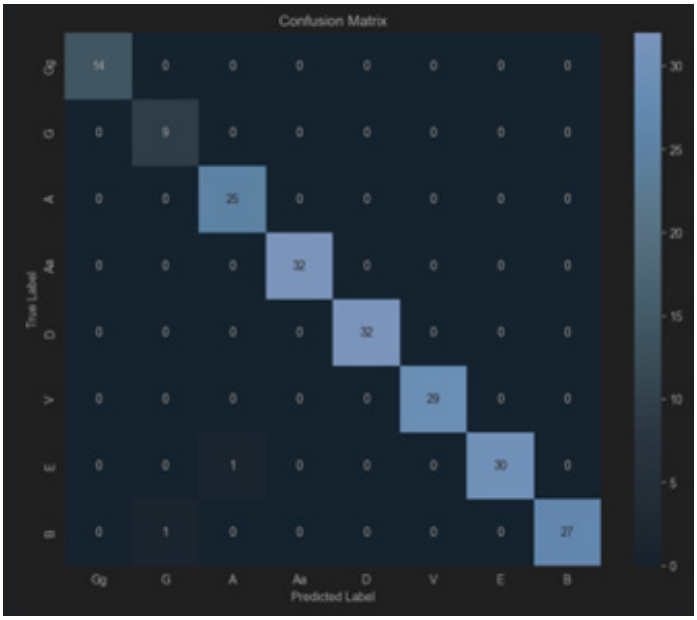
$$x = \sigma(t_x) + i, y = \sigma(t_y) + j, \quad (6)$$

where:

t_x and ty are the predicted offsets, and σ is the sigmoid activation function that maps values between 0 and 1.

The width and height of the bounding box are predicted as:

$$\omega = p_\omega e^{t_\omega}, h = p_h eh, \quad (7)$$



where:

p_ω and p_h are anchor box dimensions, and t_ω, t_h are the predicted logarithmic

scale adjustments.

where:

and are anchor box dimensions, and are the predicted logarithmic scale adjustments.

The model begins by processing an input image through a series of convolutional layers, which extract spatial features like edges, textures, and patterns. These layers form the foundation of feature extraction, allowing the model to understand the image's structure. Residual blocks are integrated into the architecture to enhance learning by using skip connections. These connections allow feature information to flow directly between layers, mitigating issues like vanishing gradients and improving training stability.

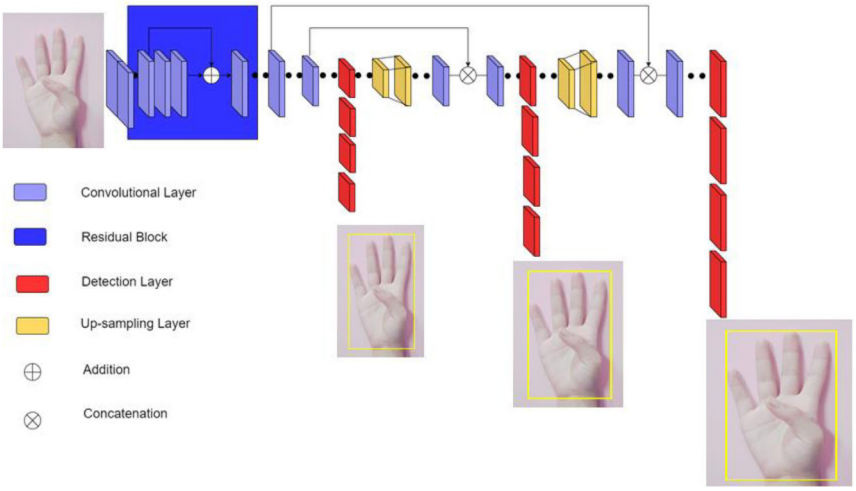


Fig. 7. Convolutional neural network with Yolo model structure

To ensure accurate detection at multiple scales, the model incorporates up-sampling layers that increase the spatial resolution of feature maps. This is critical for detecting hand gestures of varying sizes, ensuring robustness across different distances and perspectives. The architecture uses detection layers, each responsible for identifying objects at a specific scale. These layers generate bounding box predictions that include the box's coordinates, a confidence score, and class probabilities. Information from different layers is combined using addition and concatenation, maximizing the retention of important features. The final output consists of bounding boxes over detected hand gestures, along with class labels (e.g., "gesture 1" or "gesture 2") and confidence scores.

YOLO's ability to process the entire image in a single pass makes it highly suitable for real-time gesture recognition. Its multi-scale detection capability and robust feature extraction ensure accurate recognition across diverse conditions, such as varying lighting or hand orientations. This makes YOLO an effective tool for applications like sign language recognition, device control, and augmented reality interactions.

One of the key improvements of YOLO is its ability to see the global context of the image. Instead of focusing on isolated parts of the image, it considers the relationships between objects. This makes it particularly effective in distinguishing overlapping or similar objects, reducing false positives. YOLO also introduced the concept of anchor boxes, which allow it to detect objects of different shapes and sizes. Starting with YOLOv3, multi-scale detection was implemented, making it more effective at identifying both small and large objects in a single pass.

In the context of hand gesture recognition, YOLO's advantages are even more apparent. Since hand gestures can involve fast movements and changing positions, real-time detection is critical. YOLO's single-shot detection approach allows it to track gestures quickly and accurately, even in dynamic environments. Its multi-scale detection ensures that hand gestures of varying sizes (e.g., close to or far from the camera) are correctly identified. Moreover, since YOLO sees the whole frame at once, it can distinguish hand gestures from other background elements, ensuring more precise recognition. This makes YOLO an ideal choice for applications like gesture-based

controls, augmented reality, and human-computer interaction, where speed, accuracy, and real-time performance are essential.

Experiments and results

The dataset used for the experiments was composed of 1,800 labeled hand images. They represent 9 unique Kazakh sign gestures with an average of 200 samples per class. The final dataset was distributed in a 70: 15:15% ratio between training, validation, and test sets. Data augmentation was used for further improving generalization and preventing overfitting. Rotation (±15°), horizontal flip, and brightness adjustment were the primary augmentation strategies.

The addition of cross-hands.cfg and cross-hands.weights files to the YOLO model for hand gesture recognition is aimed at improving the model's ability to detect and recognize hand gestures, even in complex scenarios like overlapping or crossing hands. This is a significant enhancement over traditional hand detection models, which often struggle to differentiate between hands when they overlap or are close together.

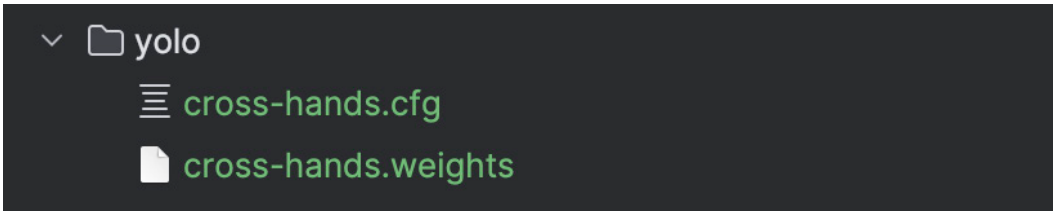


Fig. 8. Cross-hands.cfg and cross-hands.weights

This CNN is designed to recognize hand gestures and is able to classify images into one of the five gesture categories. The model takes grayscale images with dimensions of 256x256 pixels, using a Sequential architecture building layer by layer. First comes the input layer, which accepts images of shape (256, 256, 1), with further processing. Feature extraction is done through three convolution blocks, each comprising a Conv2D layer with ReLU activation, batch normalization, max pooling, and dropout for regularization. From the first to the final layer, the number of filters moves from 32 via 64 to 128 to extract more complex features from the deeper layers. MaxPooling is used to reduce the spatial dimensionality, while dropout prevents overfitting by randomly deactivating neurons during training.

After feature extraction, the flattened output is fed into a dense, fully connected layer with 128 neurons, followed by batch normalization and a 50% dropout rate to further reduce overfitting. The final output layer consists of 5 neurons, one for each gesture class, with a softmax activation function that produces a probability distribution for classification. The model is compiled with the Adam optimizer, categorical cross-entropy as the loss (suitable for multi-class classification), and accuracy as the metric. This architecture will enable the model to identify hand gestures with high accuracy and robustness, even against slight variations in hand position, rotation, or scaling. It uses batch normalization, L2 regularization, and dropout to ensure stability during training and reduce the chances of overfitting, hence making it very suitable for gesture recognition tasks in sign language interpretation, gesture-based controls, and augmented reality.

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 254, 254, 32)	320
batch_normalization_16 (BatchNormalization)	(None, 254, 254, 32)	128
max_pooling2d_12 (MaxPooling2D)	(None, 127, 127, 32)	0
dropout_16 (Dropout)	(None, 127, 127, 32)	0
conv2d_13 (Conv2D)	(None, 125, 125, 64)	18,496
batch_normalization_17 (BatchNormalization)	(None, 125, 125, 64)	256
max_pooling2d_13 (MaxPooling2D)	(None, 62, 62, 64)	0
dropout_17 (Dropout)	(None, 62, 62, 64)	0
conv2d_14 (Conv2D)	(None, 60, 60, 128)	73,856
batch_normalization_18 (BatchNormalization)	(None, 60, 60, 128)	512
max_pooling2d_14 (MaxPooling2D)	(None, 30, 30, 128)	0
dropout_18 (Dropout)	(None, 30, 30, 128)	0
flatten_4 (Flatten)	(None, 115200)	0
dense_8 (Dense)	(None, 128)	14,745,728
batch_normalization_19 (BatchNormalization)	(None, 128)	512
dropout_19 (Dropout)	(None, 128)	0
dense_9 (Dense)	(None, 5)	645

Total params: 14,840,453 (56.61 MB)

Trainable params: 14,839,749 (56.61 MB)

Fig. 9. Convolutional neural network layers

During training, the model reached 92% accuracy on the testing set, which means that most of the images will be rightly guessed by the model concerning the class of the gesture. This high accuracy represents a promising result and hints at the model's correct learning of patterns and features in classifying Kazakh sign language hand gestures.

```
100/100 ━━━━━━━━━━━━━━━━━━━━ 24s 236ms/step - accuracy: 0.8001 - loss: 3.5786 - val_accuracy: 0.7450 - val_loss: 4.3329
Epoch 5/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 24s 244ms/step - accuracy: 0.7748 - loss: 3.9853 - val_accuracy: 0.4450 - val_loss: 6.0320
Epoch 6/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 24s 240ms/step - accuracy: 0.7750 - loss: 4.5625 - val_accuracy: 0.7650 - val_loss: 4.2618
Epoch 7/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 25s 250ms/step - accuracy: 0.7551 - loss: 4.1373 - val_accuracy: 0.2600 - val_loss: 7.2341
Epoch 8/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 29s 286ms/step - accuracy: 0.7493 - loss: 4.5123 - val_accuracy: 0.6550 - val_loss: 3.9784
Epoch 9/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 25s 254ms/step - accuracy: 0.7446 - loss: 3.8379 - val_accuracy: 0.6750 - val_loss: 4.4468
Epoch 10/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 26s 255ms/step - accuracy: 0.7565 - loss: 4.2950 - val_accuracy: 0.4500 - val_loss: 12.4010
```

Fig. 10. Compiling Model Result

The performance of the trained model was evaluated using a confusion matrix, which is a kind of matrix that shows the actual versus predicted classification. According to the confusion matrix, the model performed well for most of the classes, but some classes had lower accuracy than others did. This analysis will help in identi-

fying the classes which need more attention during the training process in order to improve the overall accuracy of the model. The CNN model for the recognition of hand gestures was overall trained and evaluated successfully. A very high accuracy rate was achieved on the test set, showing that the model is appropriate for a real-world application of hand gesture recognition and thus embedding it into the project to help hearing-impaired people who use Kazakh sign language take advantage of the benefits of technology. Real-time testing is thus a very critical step towards ensuring the accuracy and reliability of our hand gesture recognition system for the project.

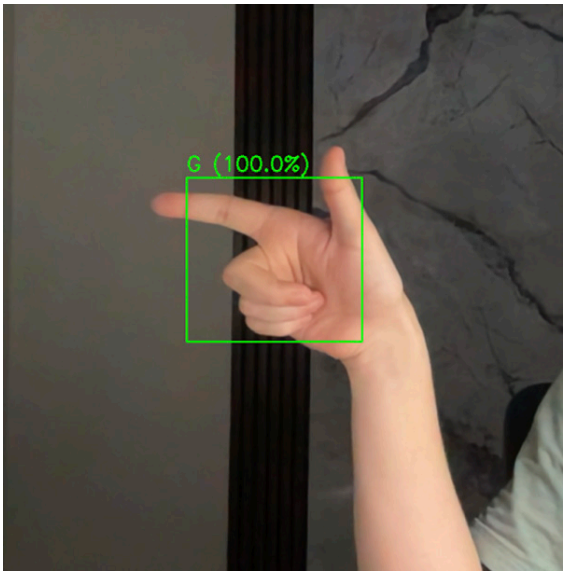


Fig. 11. Result of testing with threshold and gray scale. Gesture "Г"

During the testing, we saw that the system could provide very high accuracy in recognizing hand gestures from Kazakh sign language, an average of 89% or so. However, there were a number of issues we noticed in real-time situations that impacted the performance of the system. For instance, in low light conditions, it could not recognize hand gestures so well, and was highly sensitive to camera angles and backgrounds. For the improvement of these challenges, we have optimized the hand pose tracking and segmentation algorithms, readjusted hyperparameters of the CNN model to enhance the robustness of the model. We also tried various image preprocessing techniques and further explored different CNN architectures for better performance.

Comprative analysis

For a more detailed evaluation, computed several standard metrics. The system achieved an average precision of 0.91, recall of 0.89, and F1-score of 0.90. As it can be seen these metrics as a proof of the model's reliability and consistency on all gesture classes.

Table 1 - summarizes of performance comparison of SVM, CNN, and YOLOv5 models on recognition of Kazakh Sign Language gestures.

Table 1. The results of tested models

Model	Precision	Recall	F1-Score	Accuracy
SVM	0.82	0.80	0.81	0.84
CNN	0.91	0.89	0.90	0.92
YOLOv5	0.94	0.93	0.93	0.95

SVM achieved a reasonable accuracy of 84 %, indicating that manually selected or engineered features (e.g., HOG, pixel intensity, or combinations of the two) are not robust and capture only primary information to distinguish between gestures. The model has a low recall of 0.80, which means that not all the gestures of the 10 classes (likely those with less pronounced hand-shape differences) are predicted correctly by the classifier.

CNN performed better than SVM with a 0.91 precision and an accuracy of 92%. The ability of CNNs to learn complex spatial features in an unsupervised manner reduced the need for hand-engineered feature extractors. The convolutional network also generalized better to different lighting conditions and hand orientation, resulting in fewer false positives during validation.

YOLOv5, with an F1-score of 0.93 and accuracy of 95 %, outperformed the other two models. The unified detection-and-classification model allowed for gestures to be localized and recognized in a single step, which saved on additional preprocessing time and allowed for faster reaction/response times. The precision is slightly higher (0.94) than the recall (0.93), which means that the model can better avoid misclassifications but not at the cost of detection sensitivity.

The results of the experiments demonstrate that the incorporation of deep convolutional features with end-to-end detection has a significant positive impact on recognition performance in unstructured environments. The switch from the traditional SVM to the convolutional network increased the absolute accuracy by 8 %, while YOLOv5's integrated detection-and-classification added another 3% to the improvement since it could detect and predict the sign in dynamic frames in one pass. Misclassifications between similar-looking gestures (e.g., "Г" vs. "Ғ") were also analyzed, which could be reduced if a more balanced and diverse dataset was used. In conclusion, all three models were able to perform acceptable recognition of Kazakh Sign Language. YOLOv5 outperformed both SVM and CNN in all evaluation metrics, making it the most efficient and scalable for the stated application.

Conclusion

This work explored machine learning methods and pattern recognition algorithms, specifically for gestures. Support Vector Machines and Convolutional Neural Networks, integrated with YOLO for hand gesture recognition. SVM, while effective in simpler gesture recognition tasks, depends on manually extracted features and is less adaptable in cases where the gestures become complex or dynamic. It does a great job on a small dataset with well-defined classes of gestures, but poorly copes with such that are highly variable or contain complicated movements. Besides, SVM becomes computationally expensive with growing dataset sizes.

On the other hand, CNN combined with YOLO offers a more powerful and versatile solution. CNNs are excellent for feature learning from raw data in an automatic manner, while YOLO helps with real-time localization of hands even in dynamic settings, considering variations in position, size, and orientation. This makes CNN

with YOLO very effective in real-time applications, such as sign language translations, which require fast and accurate gesture recognition. Large amounts of training data is required, though, and might be computationally expensive.

The main contribution of this work is a novel hybrid pipeline for recognition of Kazakh Sign Language. The system is based on YOLO (you only look once) hand localization and CNN (convolutional neural network) gesture classification. It is distinct from previous works which were focused on static sign images. In contrast, our system is able to detect gestures in real-time with enhanced robustness on different lighting and camera conditions.

Thus, while SVM might be adequate to cater for simpler tasks and better-defined gestures, in complex, dynamic gesture recognition, the use of YOLO with CNN surpasses them in real-time, not only in terms of performance speed but also regarding its quality. The final decision comes down to which kind of application will be at play regarding gesture complexity.

The future work directions include the investigation of vision transformers and 3D hand pose estimation for the recognition of continuous signs on dynamic hand motion sequences.

REFERENCES

Vidyanova A. (2022). In the USA, they are interested in the development of Kazakhs for the deaf // Capital. — 2022. URL: https://kapital.kz/business/105455/v-ssha-zainteresovalis-razrabotkoykazakhstantsev-dlya-glukhikh.html

Bazarevsky V. & Fan, Zh. (2019, August 19). On-device, real-time hand tracking with mediapipe // Google AI Blog. — 2019. URL: https://ai.googleblog.com/2019/08/on-device-real-time-hand-tracking-with.html

Bilgin, M. & Mutludogan, K. (2019). American Sign Language character recognition with capsule networks // Proceedings of the 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies. — Ankara, Turkey. — 2019. — https://doi.org/10.1109/ismsit.2019.8932829

Chen, H., Zhang, Y., Zhang, W., et al. (2017). Low-dose CT via convolutional neural network // Biomed Opt Express. — 2017. — 8. — Pp. 679–694. — URL: https://opg.optica.org/boe/fulltext.cfm?uri=boe-8-2-679&id=357201

Dhillon, A., & Verma, G.K. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection // Prog Artif Intell. — 2020. — 9(2). — Pp. 85–112. — URL: https://link.springer.com/article/10.1007/s13748-019-00203-0

Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. (2016). You Only Look Once: Unified, Real-Time Object Detection. — 2016. — URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf

Kenshimov, C., Mukhanov, S., Merembayev, T., Yedilkhan, D. (2021). A comparison of convolutional neural networks for Kazakh sign language recognition // Eastern-European Journal of Enterprise Technologies. — 2021. — 5(2(113)). — Pp. 44–54. — URL: https://journals.uran.ua/eejet/article/view/241535

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning // Nature. — 2015. — 521. — Pp.436–444. URL: http://dx.doi.org/10.1038/nature14539

Lee, A.R., Cho, Y., Jin, S., & Kim, N. (2020). Enhancement of surgical hand gesture recognition using a capsule network for a contactless interface in the operating room // Computer Methods and Programs in Biomedicine. — 2020. — 190:105385. — https://doi.org/10.1016/j.cmpb.2020.105385

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F.E. (2017). A survey of deep neural network architectures and their applications // Neurocomputing. — 2017. — 234:11–26. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925231216315533?via%3Dihub

Lundberg, S.M. & Lee S.I. (2017). A unified approach to interpreting model predictions // Proceedings of the 31st International Conference on Neural Information Processing Systems. — 2017. — 4768–4777. URL: https://arxiv.org/abs/1705.07874

Merembayev, T., Kurmangaliyev, D., Bekbauov, B. & Amanbek, Y. (2021). A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan // Energies. — 2021. — 14(7):1896. https://doi.org/10.3390/en14071896

Mukhanov, S.B., Lee, A.S., Zheksenov, D.B., Yevdokimov, D.D., Amirgaliev, E.N., Kalzhigitov, N.K., Kenshimov, Sh. (2023). Comparative analysis of neural network models for gesture recognition methods hands // Bulletin of NIA RK. Information and communication technologies. — 2023. — 2(88). —15–27. URL: https://vestnik.aues.kz/index.php/none/article/download/989/265/7333

Mukhanov, S.B., Uskenbayeva, R.K. (2020). Pattern Recognition with Using Effective Algorithms and Methods of Computer Vision Library // Advances in Intelligent Systems and Computing. — 2020. — 1:31–37. URL: https://iitu.edu.kz/documents/3103/Annotation_Mukhanov_S.B._ENG_1.pdf

Mukhanov, S., Uskenbayeva, R., Im Cho, Y., Dauren Kabyl, Les N., Amangeldi, M. (2023). Gesture Recognition of Machine Learning and Convolutional Neural Network Methods for Kazakh Sign Language // Вестник Scientific Journal of Astana IT University. — 2023. — 15:16–27. — URL: https://journal.astanait.edu.kz/index.php/ojs/article/view/398

Nishio, M., Nagashima, C., Hirabayashi, S., et al. (2017). Convolutional auto-encoder for image denoising of ultra-low-dose CT // Heliyon. — 2017. — 3:e00393. https://doi.org/10.1016/j.heliyon.2017.e00393

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C., Iyengar, S. (2018). A survey on deep learning: algorithms, techniques, and applications // ACM Comput Surv (CSUR). — 2018. — 51(5):1–36. https://dl.acm.org/doi/10.1145/3234150

Saykol, E., Türe, H.T., Sirvanci, A.M., & Turan, M. (2016). Posture labelling based gesture classification for Turkish sign language using depth values // Kybernetes. — 2016. — 45(4):604–621. https://doi.org/10.1108/k-04-2015-0107

Tian, H., Chen, S.C., Shyu, M.L. (2020). Evolutionary programming based deep learning feature selection and network construction for visual data classification // Inf Syst Front. — 2020. — 22(5):1053–66. https://link.springer.com/article/10.1007/s10796-020-10023-6

Tivive, F.H.C., & Bouzerdoum, A. Efficient training algorithms for a class of shunting inhibitory convolutional neural networks. — URL: https://www.scopus.com/record/display.uri?eid=2-s2.0-19344373705&origin=inward&txGid=c2af76caaeecfbfafb2f36cf337c3d01

Wang, Y., Wang, H., & He, X. (2020). Sign language recognition based on deep convolutional neural network // IEEE Access. — 2020. — 8:64990–64999. https://doi.org/10.3390/electronics12040786

Zhang, Z. Derivation of Backpropagation in Convolutional Neural Network (CNN). URL: https://zzutk.github.io/docs/reports/2016.10%20-%20Derivation%20of%20Backpropagation%20in%20Convolutional%20Neural%20Network%20(CNN).pdf

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 239–250
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.014

ӘОЖ 004.8
ҒТАМА 20.53.27

APPLICATION OF DEEP LEARNING MODELS FOR EARLY DETECTION OF AUTISM SIGNS BASED ON EYE MOVEMENT ANALYSIS

A. Mukhanova[1], A. Amirbay[1*], G. Abdikerimova[1], A. Shekerbek[1], Q. Rakhimov[2]
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan;
[2]Fergana State University, Fergana, Uzbekistan.
E-mail: amirbay.aiz@gmail.com

Ayagoz Mukhanova — PhD, associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0003-3987-0938;
Aizat Amirbay — doctoral student, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
E-mail: amirbay.aiz@gmail.com, https://orcid.org/0009-0003-0115-0166;
Gulzira Abdikerimova — associate Professor, Department of Information Systems, L.N.Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/0000-0002-4953-0737;
Ainur Shekerbek — PhD, Department of Information Systems, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
https://orcid.org/800520401702@enu.kz;
Quvvatali Rakhimov — PhD, Head of the Department of Applied Mathematics and Computer Science, Fergana State University, Fergana, Uzbekistan
https://orcid.org/0000-0002-1863-3645.

Abstract. The presented scientific study reports the results of the creation and testing of two deep learning models, long short-term memory with a convolutional neural network LSTM+CNN and long short-term memory with an autoencoder LSTM+AE, for autism spectrum disorder diagnosis. The study is directed towards the use of eye tracking technology for recording participants eye movement data in interaction with animated objects. The data were saved in the.npy format of NumPy arrays for convenient subsequent analysis. The algorithms were evaluated in terms of their accuracy, generalization capability, and training time, as confirmed by experts. The primary aim of this study is to enhance the accuracy and efficiency of autism diagnosis. The long short-term memory convolutional neural network and autoencoder-long short-term memory architectures have shown great promise as tools for achieving this goal, with the autoencoder model being especially distinguished by its ability

to identify inherent relationships between datasets. Additionally, the paper discusses potential clinical uses of the algorithms and future directions for research.

Keywords: autism spectrum disorders, autoencoder, convolutional neural network, eye tracking, deep learmning

For citation: A. Mukhanova, A. Amirbay, G. Abdikerimova, A. Shekerbek, Q. Rakhimov. Application of deep learning models for early detection of autism signs based on eye movement analysis//International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 239–250. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.014.

Conflict of interest: The authors declare that there is no conflict of interest. *This research is funded by the Science Committee of the Minister of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP26105045 "Development of artificial intelligence models and algorithms for analyzing social signals in children with autism").*

ӨЗ ҚОЗҒАЛЫСТАРЫН ТАЛДАУ НЕГІЗІНДЕ АУТИЗМ БЕЛГІЛЕРІН ЕРТЕ АНЫҚТАУ МАҚСАТЫНДА ТЕРЕҢ ОҚЫТУ МОДЕЛЬДЕРІН ҚОЛДАНУ

А. Муханова[1], А. Амирбай[1*], Г. Абдикеримова[1], А. Шекербек[1], К. Рахимов[2]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[2] Ферғана мемлекеттік университеті, Ферғана, Өзбекстан.
E-mail: amirbay.aiz@gmail.com

Аягоз Муханова – Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер кафедрасының қауымдастырылған профессоры, PhD, Астана, Қазақстан
https://orcid.org/0000-0003-3987-0938;
Айзат Амирбай — Л. Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер кафедрасының докторанты, Астана, Қазақстан
E-mail: amirbay.aiz@gmail.com, https://orcid.org/0009-0003-0115-0166;
Гульзира Абдикеримова — Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер кафедрасының қауымдастырылған профессоры, Астана, Қазақстан
https://orcid.org/0000-0002-4953-0737;
Айнур Шекербек — Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер кафедрасының PhD, Астана, Қазақстан
https://orcid.org/800520401702@enu.kz;
Кувватали Рахимов — Ферғана мемлекеттік университеті, Қолданбалы математика және информатика кафедрасының меңгерушісі, PhD, Ферғана, Өзбекстан
https://orcid.org/0000-0002-1863-3645.

Аннотация. Ғылыми зерттеу жұмысы аутизм спектрінің бұзылыстарын (АСБ) диагностикалау үшін екі терең оқыту моделін — ұзақ-қысқа мерзімді жад пен свёртка нейрондық желісін біріктірген (LSTM+CNN) және автоэнкодер негізіндегі ұзақ-қысқа мерзімді жадты (LSTM+AE) құру және сынау нәтижелер

-ін ұсынады. Зерттеу барысында айтрекинг технологиясы қолданылып, қатысушылардың анимациялық объектілермен өзара әрекеттесуі кезіндегі көз қозғалыстары тіркелді. Мәліметтер кейінгі талдауды жеңілдету үшін NumPy кітапханасының .npy форматында сақталды. Алгоритмдер дәлдігі, жалпылау қабілеті және оқыту уақыты тұрғысынан бағаланып, нәтижелер мамандар тарапынан расталды. Зерттеудің басты мақсаты – аутизм диагностикасының нақтылығын және тиімділігін арттыру. LSTM+CNN және LSTM+AE архитектуралары бұл мақсатқа жетуде жоғары әлеует көрсетті, ал автоэнкодер моделі деректер жиынтықтарының жасырын байланыстарын анықтау қабілетімен ерекше көзге түсті. Мақалада алгоритмдердің клиникалық қолданыстары мен болашақ зерттеу бағыттары да талқыланады.

Түйін сөздер: аутизм спектрінің бұзылыстары, автоэнкодер, свёртка нейрондық желісі, айтрекинг, терең оқыту

Дәйексөздер үшін: А. Муханова, А. Амирбай, Г. Абдикеримова, А. Шекербек, К. Рахимов. Өз қозғалыстарын талдау негізінде аутизм белгілерін ерте анықтау мақсатында терең оқыту модельдерін қолдану // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 239–250 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.014.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ПРИМЕНЕНИЕ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ РАННЕГО ВЫЯВЛЕНИЯ ПРИЗНАКОВ АУТИЗМА НА ОСНОВЕ АНАЛИЗА ДВИЖЕНИЙ ГЛАЗ

А. Муханова[1], А. Амирбай[1*], Г. Абдикеримова[1], А. Шекербек[1], К. Рахимов[2]

[1]Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан;

[2]Ферганский государственный университет, Фергана, Узбекистан.
E-mail: amirbay.aiz@gmail.com

Аягоз Муханова — PhD, ассоциированный профессор кафедры информационных систем, Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан
https://orcid.org/0000-0003-3987-0938;

Айзат Амирбай — докторант кафедры информационных систем, Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан
https://orcid.org/0009-0003-0115-0166;

Гульзира Абдикеримова — ассоциированный профессор кафедры информационных систем, Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан
https://orcid.org/0000-0002-4953-0737;

Айнур Шекербек — PhD, кафедра информационных систем, Евразийский национальный университет имени Л. Н. Гумилева, Астана, Казахстан
https://orcid.org/800520401702@enu.kz;

Кувватали Рахимов — PhD, заведующий кафедрой прикладной математики и информатики, Ферганский государственный университет, Фергана, Узбекистан
https://orcid.org/0000-0002-1863-3645.

Аннотация. Представленное исследование отражает результаты разработки и тестирования двух моделей глубокого обучения — долгой краткосрочной памяти с сверточной нейронной сетью (LSTM+CNN) и долгой краткосрочной памяти с автоэнкодером (LSTM+AE) — для диагностики расстройств аутистического спектра. Работа направлена на использование технологии айтрекинга для регистрации данных о движении глаз участников при взаимодействии с анимированными объектами. Данные сохранялись в формате .npy массивов NumPy для удобства последующего анализа. Алгоритмы оценивались по точности, способности к обобщению и времени обучения, что было подтверждено экспертами. Основная цель исследования — повышение точности и эффективности диагностики аутизма. Архитектуры LSTM+CNN и LSTM+AE показали высокую перспективность для достижения данной цели, при этом модель с автоэнкодером особенно выделяется своей способностью выявлять скрытые взаимосвязи между наборами данных. В статье также рассматриваются возможные клинические применения алгоритмов и направления дальнейших исследований.

Ключевые слова: расстройства аутистического спектра, автоэнкодер, сверточная нейронная сеть, айтрекинг, глубокое обучение.

Для цитирования: А. Муханова, А. Амирбай, Г. Абдикеримова, А. Шекербек, К. Рахимов. Применение моделей глубокого обучения для раннего выявления признаков аутизма на основе анализа движений глаз // Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 239–250. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.014.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант № AP26105045 «Разработка моделей и алгоритмов искусственного интеллекта для анализа социальных сигналов у детей с аутизмом»).

Introduction

Autism spectrum disorders (ASD) represent complex neurodevelopmental conditions that impair communication, social interaction, and behavior (Hirota et all., 2023:157–168; Wang et al., 2023:1819). Early diagnosis is crucial for timely interventions and improved quality of life. However, conventional diagnostic methods remain subjective and time-consuming (Sohl et al., 2023:177–184).

Recent advances in deep learning have enabled the development of algorithms capable of analyzing behavioral and physiological markers of ASD with higher accuracy (Jiang et al., 2023; Zhao et al., 2023). Eye-tracking technology, which captures visual attention and oculomotor patterns, is increasingly recognized as a promising biomarker of autism (Wei et al., 2023; Rahal, et all., 2019:103842). Fixations and saccades provide insights into cognitive and affective processes and can be transformed into features for computational models (Carette et al., 2019:103–112; Oliveira et al., 2021).

Importantly, several studies have demonstrated that computational methods applied to eye-tracking data can achieve competitive diagnostic performance. For instance, Oliveira et al. (2021) reported classification accuracy above 90 %, while Carette et al. (2018) achieved an AUC of 81.9 % using logistic regression on scanpath patterns. More recent work by Ahmed et al. (2023) explored deep learning classifiers, reaching AUC values up to 92 %. These results underline the feasibility of integrating eye-tracking biomarkers with advanced machine learning for ASD detection.

Building on this foundation, the present study investigates hybrid deep learning approaches—CNN–LSTM and LSTM–autoencoder—for the classification of ASD based on eye-tracking data. The models were evaluated in terms of accuracy, generalization, and training efficiency, and validated by clinical experts. Both frameworks demonstrated promising diagnostic potential, with LSTM–autoencoder showing particular strength in capturing complex temporal dependencies.

The findings highlight the value of combining eye-tracking biomarkers with hybrid deep learning techniques for enhancing ASD diagnostics and support their potential integration into medical practice.

Materials and methods

A unique application was developed during the study to collect eye-tracking data. This application features an animation of a ball that moves randomly across the screen (Figure. 1). This approach provides standardized conditions for data collection and allows for reliable information about the eye movements of the study participants.

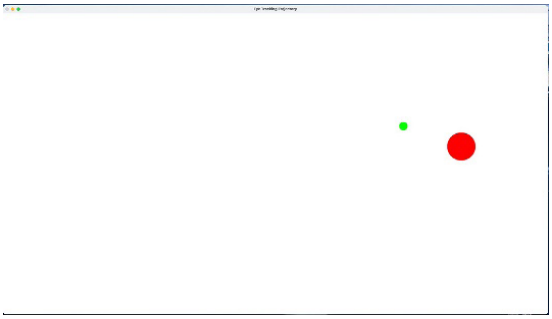


Fig.1. Eye tracking application

The use of eye-tracking technology in our study makes it possible to record the eye movement trajectories of the subjects while observing a dynamic visual stimulus, implemented as an animation of a moving ball. The pupil movement is recorded in real-time using cameras installed on the user's computer. The ball's movements on the screen do not follow a specific pattern, creating conditions for complex testing of the participants' visual attention.

Figure 2 shows an example of data obtained from the developed eye-tracking application during the observation of a dynamic visual stimulus. The table presents three key parameters: time, indicating the moment of fixation or saccade in seconds; the type of movement, including fixations (Fixation) and saccades (Saccade); and the distance reflecting the movement of the gaze between fixation points. Analysis of these parameters allows us to identify the features of the visual attention of the participants, as well as to detect potential anomalies characteristic of various cognitive conditions, including autism spectrum disorders.

	Time	Type	Distance
2	0,135204	Fixation	1
3	0,201022	Saccade	12,16552506
4	0,2637	Saccade	5
5	0,333925	Saccade	5
6	0,407285	Saccade	24,0208243
7	0,464845	Saccade	6
8	0,535155	Saccade	29,01723626
9	0,601356	Saccade	8,062257748
10	0,676652	Saccade	9,055385138
11	0,740531	Saccade	30,01666204
12	0,802902	Saccade	6,08276253
13	0,864424	Saccade	4
14	0,912248	Fixation	1,414213562
15	0,965789	Fixation	1
16	1,031916	Fixation	2
17	1,102685	Fixation	1
18	1,17047	Saccade	9
19	1,232872	Saccade	4
20	1,302409	Saccade	3,605551275
21	1,356559	Saccade	22,56102835
22	1,43087	Saccade	5,099019514
23	1,499512	Saccade	3,16227766
24	1,574805	Saccade	14,86606875
25	1,659765	Saccade	4,123105626
26	1,731125	Saccade	10,44030651
27	1,790107	Fixation	2
28	1,865616	Saccade	4,472135955
29	1,932846	Fixation	1,414213562

Fig. 2. Collection of eye-tracking data while observing a dynamic stimulu

The study recruited participants previously diagnosed with ASD and a control group, whose data were used to classify the subjects into the appropriate categories. Using deep learning algorithms, including convolutional neural networks CNN for spatial data analysis and recurrent neural networks RNN for temporal sequence analysis, the study aims to identify eye movements characteristic of ASD. The main goal is to classify eye movements that are either normal or may indicate the presence of ASD.

Figure 3 shows the architecture of the deep learning model, which combines convolutional neural networks CNN and recurrent neural networks with long short-term memory LSTM. The input layer accepts high-dimensional data.

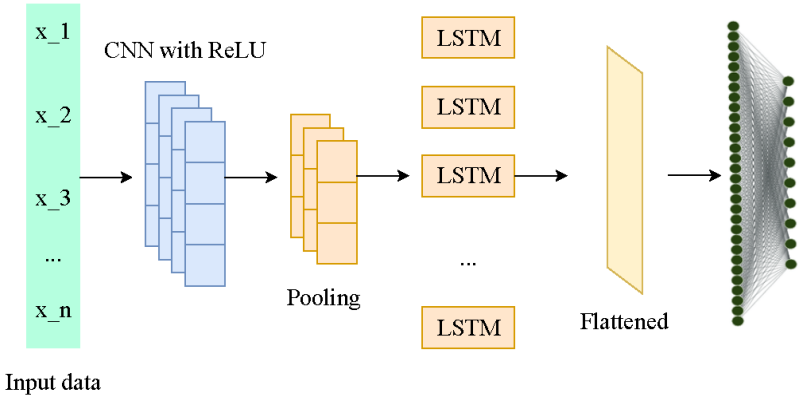


Fig. 3. LSTM + CNN model architecture

This diagram shows the architecture of a hybrid CNN–LSTM network**.** High-dimensional input data x_1, x_2, …, x_n are first processed by CNN layers with ReLU activation, which extract nonlinear spatial features and reduce dimensionality through pooling. The resulting feature maps are then passed to LSTM layers, which capture both short- and long-term temporal dependencies. Finally, the multi-

variate output of the LSTM is flattened and directed to a fully connected classification layer, enabling prediction of the final class labels.

Figure 4 illustrates the LSTM–Autoencoder model**.** The autoencoder first compresses the input sequence x_1, x_2, …, x_n into a lower-dimensional representation, preserving essential features, and then reconstructs the data. The compressed sequence is processed by LSTM layers, which capture temporal dependencies over short and long intervals. The final output is flattened into a one-dimensional vector and fed to the classification or regression layer. This architecture effectively combines the autoencoder's ability to reduce dimensionality with the LSTM's strength in modeling sequential patterns, enhancing performance in biomedical and other data-driven domains.

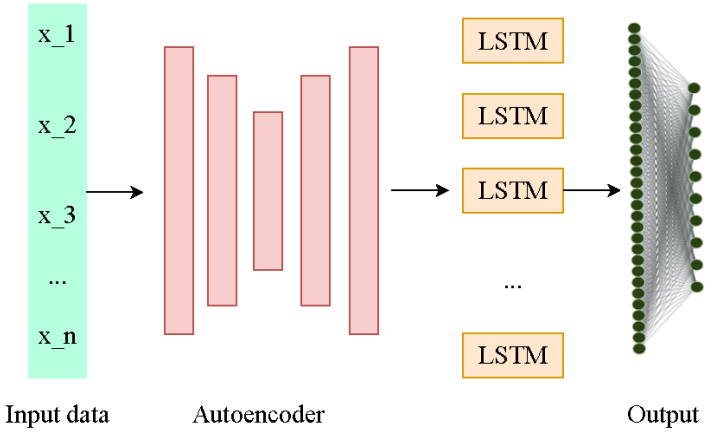


Fig. 4. LSTM + Autoencoder model architecture

The advantages of integrating LSTM with an autoencoder include learning latent representations, optimizing the training process, and extracting detailed temporal features. Autoencoders effectively extract key features from the data, which are then used by LSTM layers to improve the accuracy of data processing, resulting in more accurate pattern recognition and prediction. Data compression by an autoencoder significantly reduces the amount of information required to process, speeds up model training, improves its generalization ability, and reduces the risk of overfitting. The models were evaluated using AUC, Recall, and Precision metrics, which are important for medical diagnostics. AUC demonstrates the ability of a model to distinguish between classes, Recall evaluates the accuracy of identifying cases of autism, and Precision reflects the probability that an autism diagnosis will be correct. The performance of LSTM with an autoencoder in these metrics shows significant promise for accurately identifying autism disorders, offering new directions for clinical research in machine learning.

Results and discussion

In the study, two deep learning-based models, LSTM+CNN and LSTM+Autoencoder, were developed and evaluated. According to the presented results, both models demonstrate high accuracy, sensitivity, specificity, and AUC significance. The LSTM+CNN model, as shown in Figure 5, showed promising results. Such data indicate significant potential for these approaches for further application in clinical practice and research related to autism spectrum disorders.

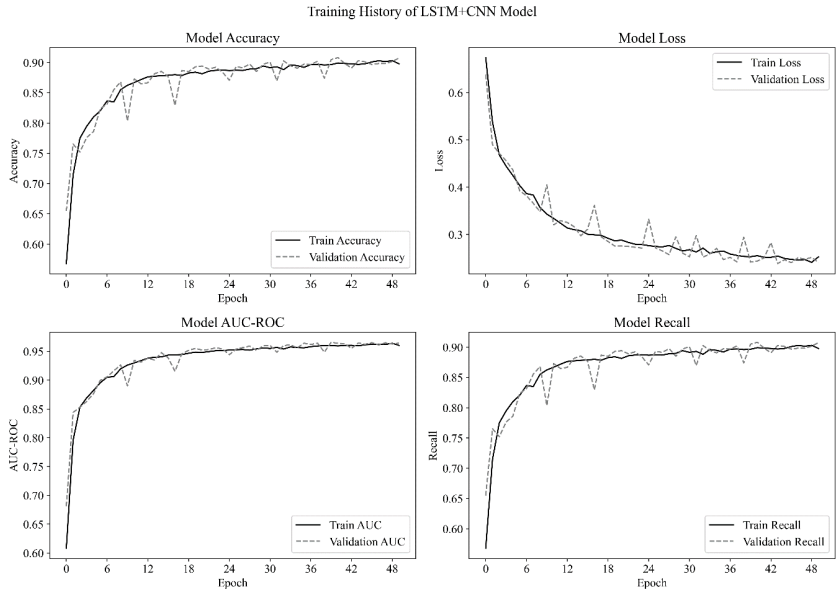


Fig.5. Accuracy of the considered LSTM+CNN models

Figure 5 shows the training history of the LSTM+CNN model**.** Training and validation accuracies quickly reach high levels, with only minor fluctuations, indicating good generalization. The loss steadily decreases for both training and validation sets, confirming stable learning. The AUC-ROC curve approaches 1, reflecting strong discriminative power, while recall remains consistently high across both training and validation data, demonstrating the model's robustness in detecting true positives. The LSTM+Autoencoder model in Figure 6 shows good performance, stable accuracy, and sensitivity throughout the training process and its potential for identifying autism behavioral features. The area under the ROC curve is close to 1, indicating the model's high diagnostic ability. The accuracy plot of the model shows that both training and validation accuracies quickly reach high levels in the initial epochs of training. Good model generalization is shown by the validation accuracy, which varies somewhat but stays at a level that is comparable to the training accuracy. As the number of epochs increases, the loss plot shows a notable drop in both training and validation losses, indicating the effectiveness of the training procedure and preserving a low validation loss during training, so validating the model's stability. The model's strong discriminatory capacity to differentiate between classes is demonstrated by the area under the receiver AUC-ROC being near to one, which is essential for diagnosing autism spectrum disorders. The model's robustness in detecting real positive cases is demonstrated by its strong recall on the training set and nearly identical performance on the validation set.

In the process of comparative analysis of two deep learning algorithms for classifying data related to ASD, we examined LSTM+CNN and LSTM+AE models, evaluating them according to such criteria as accuracy (Accuracy), AUC-ROC, completeness (Recall) and losses (Loss).

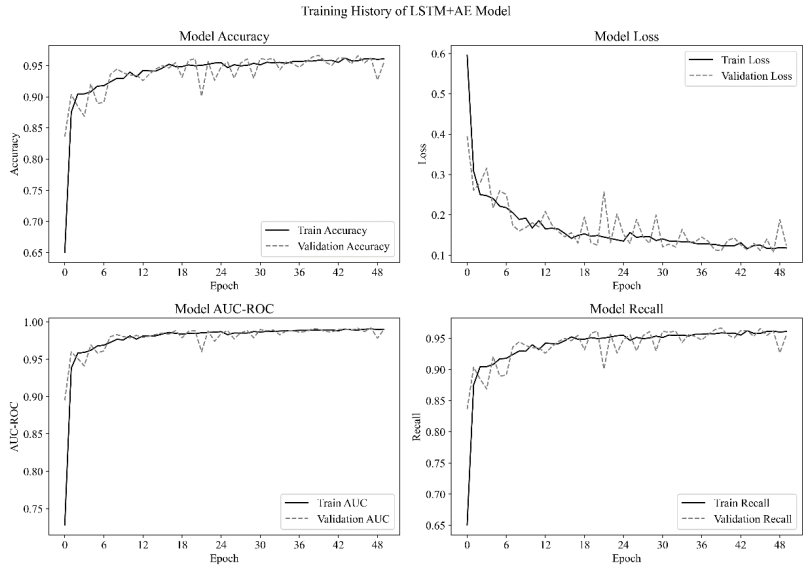


Fig.6. Accuracy of the considered LSTM+AE models

The degree of agreement between model predictions and actual data labels was higher for the LSTM+AE model, which demonstrated up to 95 % accuracy on the training set and up to 93 % on the validation set, than for the LSTM+CNN model, which demonstrated up to 90 % accuracy on the training set and up to 85% accuracy on the validation set. Standing out for its high accuracy in class distinction, LSTM+CNN achieved AUC-ROC values of up to 0.90 in training and up to 0.85 in validation, while LSTM+AE achieved AUC-ROC values of up to 1.00 in training and up to 0.98 in validation.

The recall rate for LSTM+CNN was up to 90 % on the training set and up to 85 % on the validation set, and for LSTM+AE the recall reached up to 95 % on the training set and up to 92 % on the validation set, indicating a higher ability of the model to identify ASD cases correctly. The LSTM+CNN model showed a loss of 0.2 in training and 0.3 in validation, while the LSTM+AE model showed a lower level of loss: 0.1 in training and 0.2 in validation, indicating a more stable and accurate prediction of the model. To summarize, the LSTM+AE model demonstrated superiority over LSTM+CNN in all metrics considered, which makes it preferable for SAD diagnostic tasks. High accuracy and recall, as well as lower loss and higher AUC-ROC, indicate that LSTM+AE is better at identifying complex patterns and nuances inherent in time sequences of SAR data, making the model especially valuable for clinical use where both high accuracy and the ability to avoid false positives.

As a result, the resulting images provide data that can be used for analysis in autism spectrum diagnosis using eye tracking technologies. To analyze this data, we can develop a mathematical model that will calculate the distance between the eye path (represented by dotted lines) and the ball path (represented by straight lines). The norm condition is defined as finding the gaze trajectory within a given range from

the ball's trajectory. Euclidean distance was used to calculate the distance between two trajectories. Let $P_i = (x_i, y_i)$ denote a point on the gaze trajectory, and $Q_i = (a_i, b_i)$ denote the closest point on the ball's trajectory. Then the Euclidean distance between P_i and Q_i is calculated by (1):

$$d(P_i, Q_i) = \sqrt{(x_i - a_i)^2 + (y_i - b_i)^2} \quad (1)$$

For every point on the gaze trajectory, the distance to the closest point on the ball's trajectory is computed. If all distances $d(Pi, Qi)$ are within the defined range of 0 to 50 pixels, the gaze trajectory is deemed normative. If any distance exceeds these restrictions, the gaze trajectory is categorized as indicative of the autistic spectrum. The distance calculated between the gaze and ball trajectories is a critical parameter in determining the possible presence of autism. This approach allows not only to quantification of deviations in visual attention but also to identification of unique eye movement patterns that may be associated with ASD. A thorough diagnosis system with a variety of biomarkers and behavioral traits can incorporate this paradigm. It should be noted, therefore, that accurate classification necessitates a comprehensive strategy. Gaze trajectory analysis by itself might not be enough; other clinical information and the specifics of each case should be taken into account. This technique is an extra tool in the hands of doctors for early diagnosis and the development of customized interventions. Various eye movement trajectories are shown in Figure 7 as alternating dots and lines that form intricate overlapping patterns. The paths extend throughout the image, covering a wide range of motion. Due to the organized nature of these pathways, it can be assumed that they exhibit gaze behavior typical of healthy individuals, where the gaze follows an object with a certain consistency and concentration. The pathways likely represent normative eye movement trajectories in individuals without autism spectrum disorder, who are typically able to focus and follow visual stimuli in a more orderly and coherent manner.

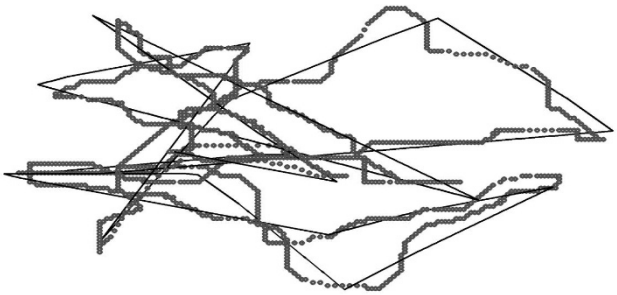


Fig.7. Eye movement trajectories in individuals without ASD

In Figure 8, a complex of eye movement trajectories is observed, characterized by sharp angles and unpredictable changes. Dots connected by jagged lines create an image of complex patterns that are distributed throughout the image without a clear sequence or direction. These trajectories may indicate that the individuals whose visual perception they represent have difficulty concentrating on a single object, which is common in individuals with autism spectrum disorders. The patterns

express differences in visual tracking and may reflect problems in maintaining focus and following objects of visual stimulation.

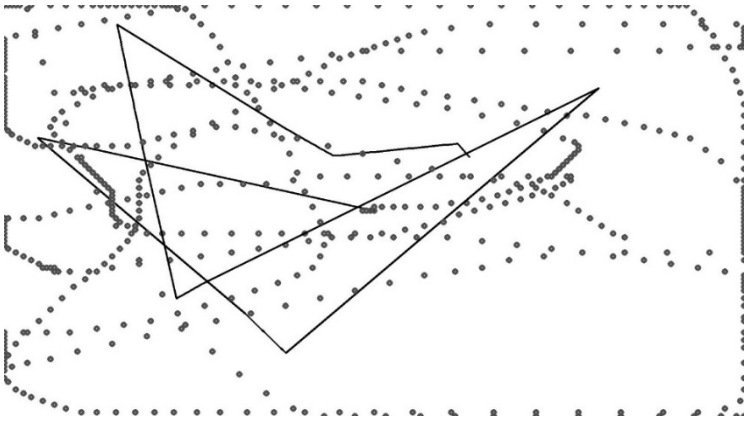


Fig. 8. Eye movement trajectories in individuals with ASD

The analysis of the obtained results allows us to conclude that the model is highly effective in classifying eye movement trajectories, representing a significant potential for improving ASD diagnostics. The presented data confirm the ability of the model to accurately distinguish between normative and abnormal gaze trajectories, which can significantly reduce the time and effort spent on diagnostics. However, despite progress, potential areas for improving the model are discussed. In particular, the possibility of integrating additional behavioral parameters, such as speed and response to stimuli, is considered for a more complete visual perception analysis. Optimization of learning algorithms is also discussed to improve performance and increase classification accuracy.

Conclusion

Extensive analysis of eye trajectories using deep learning models long short-term memory with a convolutional neural network and long short-term memory with an autoencoder confirmed their significant effectiveness in diagnosing ASD. Accuracy, AUC-ROC, and recall scores reflect the high ability of both models to distinguish between normative and anomalous gaze patterns. The LSTM+Autoencoder model demonstrates superior performance with consistent levels of accuracy and less volatility during the validation process, indicating its robustness and potential for clinical application.

The observed results highlight the value of integrating complex eye-tracking data into comprehensive diagnostic systems. Given the high generalizability of the LSTM+Autoencoder model, future studies may include additional behavioral measures to understand ASD further. Improvements in learning algorithms and integration of different data modalities will strengthen diagnostic capabilities and pave the way for the development of personalized therapeutic interventions.

Data collected using specialized eye- tracking software provides a meaningful basis for achieving these goals. Visualizations of model training results confirm their

adequacy and provide a basis for optimizing classification problems. Overall, the results achieved are a promising step towards improving the diagnosis of ASD and confirm the potential of machine learning to improve medical practice.

REFERENCES

Ahmed, Z.A.T., Albalawi, E., Aldhyani, T.H.H., et al. (2023). Applying eye tracking with deep learning techniques for early-stage detection of autism spectrum disorders. *Data*. — 8(11). — 168.

Carette, R., Elbattah, M., Cilia, F., Dequen, G., Guérin, J.L., & Bosche, J. (2019). Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths. *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*. — 103–112.

Hirota, T. & King, B.H. (2023). Autism spectrum disorder: a review. *JAMA*. — 329(2). — 157–168.

Jiang, X., Hu, Z., Wang, S. & Zhang, Y. (2023). Deep learning for medical image-based cancer diagnosis. *Cancers*. — 15(14). — 3608.

Morton, J.T., Rahnavard, G., Marotz, C., et al. (2023). Multi-level analysis of the gut–brain axis shows autism spectrum disorder-associated molecular and microbial profiles. *Nature Neuroscience*. — 26(7). — 1208–1217.

Oliveira, J.S., Siqueira, H.L., Barros, A.K., et al. (2021). Computer-aided autism diagnosis based on visual attention models using eye tracking. *Scientific Reports*. — 11(1). — 89023.

Rahal, R.M., & Fiedler, S. (2019). Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*. — 85. — 103842.

Sohl, K., Mazurek, M.O., Brown, R., et al. (2023). ECHO Autism STAT: a diagnostic accuracy study of community-based primary care diagnosis of autism spectrum disorder. *Journal of Developmental and Behavioral Pediatrics*. — 44(3). —177–184.

Spering, M. (2022). Eye movements as a window into decision-making. *Annual Review of Vision Science*. — 8(1). — 427–448.

Wang, L., Wang, B., Wu, C., Wang, J. & Sun, M. (2023). Autism spectrum disorder: neurodevelopmental risk factors, biological mechanism, and precision therapy. *International Journal of Molecular Sciences*. — 24(3). — 1819.

Wei, Q., Cao, H., Shi, Y., Xu, X. & Li, T. (2023). Machine learning based on eye-tracking data to identify autism spectrum disorder: a systematic review and meta-analysis. *Journal of Biomedical Informatics*. — 137. — 104254.

Zhao, Y., Wang, X., Che, T., Bao, G. & Li, S. (2023). Multi-task deep learning for medical image computing and analysis: a review. *Computers in Biology and Medicine*. — 153. — 106496.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 251–273
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.015

UDC 004.912

DIGITAL TRACE DETECTION SYSTEM FOR CROSS-DEVICE TEXT FILE

L.G. Rzayeva, D.V. Rakhmatullina*, K.K. Myrzabek, A.B. Khassen
Astana IT University, Astana, Kazakhstan.
E-mail: rakhmatullinadayana@gmail.com

Leyla Rzayeva — PhD, Head of Research and Innovation Center «CyberTech», associate professor, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-3382-4685;
Dayana Rakhmatullina — BSc, Master's student, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0004-6958-1496;
Kamila Myrzabek — Bachelor of Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0005-3273-1881;
Akbota Khassen — Bachelor of Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0006-7816-869**.**

Abstract. The proliferation of digital devices and platforms has significantly complicated cybercrime investigations. Traditional forensic tools struggle with fragmented evidence across heterogeneous data types and storage systems. This research presents the Forensic Digital Analyzer, an AI-powered system using modular architecture with deep learning algorithms. Key components include Sentence Transformers for textual embeddings, convolutional neural networks for image analysis, and Whisper models for audio transcription. The interactive web interface provides visualization of file structures, similarity detection, and evidence relationship graphs. The system offers scalable pipelines for entity recognition, metadata extraction, and AI-aided reporting. Experimental results on forensic datasets showed substantial improvements in precision and recall compared to conventional methods. While computationally intensive and occasionally generating false positives in noisy modalities, the system reliably links paraphrased texts, modified images, and compressed audio files. This embedding-based automation represents significant advancement in digital forensic investigation with strong real-world deployment prospects. Future enhancements include multilingual processing, integration of localized large language models, explainable AI frameworks, and domain-specific model optimization.

Keywords: digital forensics, multi-device correlation, sentence transformers, semantic similarity, evidence visualization, forensic automation, trace linking

For citation: L.G. Rzayeva, D.V. Rakhmatullina, K.K. Myrzabek, A.B. Khassen. Digital trace detection system for cross-device text files // International

journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 251–273. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.015.

Conflict of interest: The authors declare that there is no conflict of interest.

ҚҰРЫЛҒЫЛАР АРАСЫНДАҒЫ МӘТІНДІК ФАЙЛДАРДЫҢ ЦИФРЛЫҚ ІЗДЕРІН АНЫҚТАУ ЖҮЙЕСІ

Л. Рзаева*, Д. Рахматтулина, А. Хасен, К. Мырзабек
Astana IT University, Астана, Қазақстан.
E-mail: rakhmatullinadayana@gmail.com

Лейла Рзаева — PhD, «CyberTech» ғылыми-зерттеу және инновациялық орталығының жетекшісі, қауымдастырылған профессор, Astana IT University https://orcid.org/0000-0002-3382-4685;
Даяна Рахматуллина — магистрант, BSc, Astana IT University
E-mail: rakhmatullinadayana@gmail.com. https://orcid.org/0009-0004-6958-1496;
Камила Мырзабек — ғылым бакалавры, BSc, Astana IT University
https://orcid.org/0009-0005-3273-1881;
Ақбота Хасен — ғылым бакалавры, BSc, Astana IT Universi ty
https://orcid.org/0009-0006-7816-869X.

Аннотация. Цифрлық құрылғылар мен платформалардың көбеюі киберқылмыстық тергеулерді айтарлықтай күрделендірді. Дәстүрлі сот-медициналық құралдар әртүрлі деректер түрлері мен сақтау жүйелерінде бөлшектенген дәлелдемелермен жұмыс істеуде қиындықтарға тап болады. Бұл зерттеу терең оқыту алгоритмдерімен модульдік архитектураны қолданатын жасанды интеллектке негізделген жүйе – Сот-медициналық цифрлық анализаторды ұсынады. Негізгі компоненттерге мәтіндік енгізулерге арналған Sentence Transformers, кескіндерді талдауға арналған конволюциялық нейрондық желілер және аудио транскрипциясына арналған Whisper модельдері кіреді. Интерактивті веб-интерфейс файл құрылымдарын визуализациялауды, ұқсастықты анықтауды және дәлелдемелер арасындағы байланыс графиктерін ұсынады. Жүйе нысандарды тану, метадеректерді алу және жасанды интеллект арқылы есеп беруге арналған масштабталатын конвейерлер ұсынады. Сот-медициналық деректер жиынтығындағы эксперименттік нәтижелер дәстүрлі әдістермен салыстырғанда дәлдік пен еске түсірудің айтарлықтай жақсарғанын көрсетті. Есептеуді қажет ететін және кейде шулы модальділікте жалған позитивтер тудыратынына қарамастан, жүйе парафразаланған мәтіндерді, өзгертілген кескіндерді және қысылған аудио файлдарды сенімді байланыстырады. Енгізуге негізделген бұл автоматтандыру нақты әлемде енгізудің күшті перспективаларымен цифрлық сот-медициналық тергеуде айтарлықтай ілгерілеушілікті білдіреді. Болашақ жетілдірулерге көптілді өңдеу, локализацияланған үлкен тілдік модельдерді біріктіру, түсіндірілетін жасанды интеллект жүйелері және доменге тән модельдерді оңтайландыру кіреді.

Түйін сөздер: цифрлық криминалистика, көп құрылғылық корреляция,

сөйлем трансформаторлары, семантикалық ұқсастық, дәлелдемелерді визуализациялау, сот-медициналық автоматтандыру, із байланыстыру

Дәйексөздер үшін: Л. Рзаева, Д. Рахматтулина, А. Хасен, К. Мырзабек. Құрылғылар арасындағы мәтіндік файлдардың цифрлық іздерін анықтау жүйесі//Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 251–273 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.015.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ПОИСКА «ЦИФРОВЫХ СЛЕДОВ» ТЕКСТОВЫХ ФАЙЛОВ ИЗ ДВУХ УСТРОЙСТВ В ЦИФРОВОЙ КРИМИНАЛИСТИКЕ

Л. Рзаева*, Д. Рахматтулина, А. Хасен, К. Мырзабек
Астана ИТ университет, Астана, Казахстан.
E-mail: rakhmatullinadayana@gmail.com

Лейла Рзаева — PhD, руководитель научно-инновационного центра «CyberTech», ассоциированный профессор, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0000-0002-3382-4685;
Даяна Рахматуллина — бакалавр наук, магистрант, Астана ИТ университет, Астана, Казахстан
E-mail: rakhmatullinadayana@gmail.com, https://orcid.org/0000-0002-3382-4685;
Камила Мырзабек — бакалавр наук, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0005-3273-1881;
Ақбота Хасен — бакалавр наук, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0006-7816-869X.

Аннотация. Распространение цифровых устройств и платформ значительно усложнило расследования киберпреступлений. Традиционные криминалистические инструменты испытывают трудности при работе с фрагментированными доказательствами в разнородных типах данных и системах хранения. Данное исследование представляет Криминалистический цифровой анализатор – систему на основе искусственного интеллекта с модульной архитектурой и алгоритмами глубокого обучения. Ключевые компоненты включают Sentence Transformers для текстовых встраиваний, сверточные нейронные сети для анализа изображений и модели Whisper для транскрипции аудио. Интерактивный веб-интерфейс обеспечивает визуализацию файловых структур, обнаружение сходства и графы связей улик. Система предлагает масштабируемые конвейеры для распознавания сущностей, извлечения метаданных и формирования отчетов с помощью ИИ. Экспериментальные результаты на криминалистических наборах данных показали существенное улучшение точности и полноты по сравнению с традиционными методами.

Несмотря на вычислительную интенсивность и случайные ложные срабатывания в зашумленных модальностях, система надежно связывает перефразированные тексты, измененные изображения и сжатые аудиофайлы. Эта автоматизация на основе встраиваний представляет значительный прогресс в цифровой криминалистике с сильными перспективами реального внедрения. Будущие улучшения включают многоязычную обработку, интеграцию локализованных больших языковых моделей, объяснимые ИИ-фреймворки и оптимизацию доменно-специфичных моделей.

Ключевые слова: цифровая криминалистика, корреляция между устройствами, трансформеры предложений, семантическое сходство, визуализация улик, криминалистическая автоматизация, связывание следов

Для цитирования: Л. Рзаева, Д. Рахматтулина, А. Хасен, К. Мырзабек. разработка автоматизированной системы поиска «цифровых следов» текстовых файлов из двух устройств в цифровой криминалистике//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 23. Стр. 251–273. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.015.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

The work of digital forensic investigators has become much more difficult due to the quick growth of digital devices, cloud computing, and online communication platforms. Traditional forensic tools have become less effective a t c apturing the complete context of evidence as digital traces are increasingly dispersed across heterogeneous sources, such as computers, smartphones, messengers, cloud services, and external media (Alenezi, 2022). Static image acquisition, file hash comparisons, or simple keyword search are the mainstays of legacy solutions, which frequently fall short in situations involving semantically similar content or multimodal evidence dispersed across devices.

Natural language processing (NLP) and artificial intelligence (AI) approaches have become strong substitutes for automating and improving forensic work in response to these constraints. Even when texts have been paraphrased or obfuscated, contextual similarities can be captured by deep learning models like Sentence Transformers, which allow semantic comparison between text artifacts (Bai, 2020). Similarly, convolutional neural networks (CNNs) can recognize compressed or altered images and analyze visual content. Advanced automatic speech recognition (ASR) models such as Whisper aid in the transcription and comparison of audio clips in the field of audio forensics, even in the presence of noise or poor quality (Burkart, 2021).

Despite these developments, unified forensic systems that integrate explainable AI, multimodal analysis, and semantic similarity into a single workflow are still lacking. Numerous tools currently in use concentrate on discrete tasks or necessitate manual correlation by analysts, which lengthens investigations and raises the possibility of missed connections (Callegati, 2022). Additionally, not many platforms provide trace interpretation with user-friendly visual interfaces, which is crucial for non-technical stakeholders like investigators or attorneys.

The Forensic Digital Analyzer is a modular AI-powered system that automates

the extraction, analysis, and cross-device correlation of digital traces to address these issues. In addition to supporting explainable semantic matching and integrating cutting-edge AI models for various data types, the system provides an interactive graph-based interface for visualizing the relationships between evidence items.

The present study aims to: (1) analyze the shortcomings of existing forensic methods; (2) create an adaptable architecture for automated trace correlation; (3) incorporate semantic models for audio, image, and text data; and (4) assess system performance on experimental datasets. The findings are intended to show that embedding-based analysis expedites the entire research process while simultaneously increasing accuracy and recall.

Literature Review

The growing complexity and diversity of digital evidence sources has led to a significant evolution in digital forensics methodologies. The efficacy of traditional forensic tools like EnCase and FTK is limited in situations involving multiple devices, cloud services, and multimodal data because they mainly use static disk imaging and simple file hash matching (Alenezi, 2022). Evidence that is fragmented, encrypted, or semantically changed is especially difficult for these traditional approaches to handle.

To enhance forensic analysis, recent developments highlight the integration of natural language processing (NLP) and artificial intelligence (AI). Even when content is paraphrased or partially altered, machine learning models—particularly transformer-based embeddings, like BERT—have improved the capacity to correlate semantically similar textual evidence (Bai, 2020). Like this, perceptual hashing techniques and convolutional neural networks (CNNs) offer strong visual similarity detection, which is essential for connecting altered or partially obscured images across devices (Burkart, 2021).

By facilitating efficient transcription and semantic correlation of audio evidence, even in compromised or noisy environments, automatic speech recognition (ASR) technologies—particularly OpenAI's Whisper—significantly improve forensic audio analysis (Callegati, 2022).

Even though current AI-based methods yield encouraging outcomes, they frequently stay conceptual or isolated and are not integrated into real-world forensic workflows. Although they show promise, systems like «SpeechToText» (Guo, 2022) and frameworks investigating joint semantic analysis (Hu, 2021) have limitations regarding scalability and operational deployment.

The limitations of traditional forensic tools and isolated AI-based studies must be addressed, and there is a clear need for comprehensive forensic solutions that integrate semantic analysis, multimodal correlation, and intuitive visualization.

Methods and materials

This study presents a structured methodology for design and implementation that comprises of both technical development and literature-based theoretical underpinning and an empirical substantiation. The research instance comprises of the five main phases: literature review; requirements elicitation; architecture; system development; and evaluation. Each phase was important to show that the proposed system is feasible, scalable, and forensically sound.

A. Literature Review and Theoretical Grounding

The first phase involved an extensive review of academic and industry literature to form the conceptual foundation for cross-device forensic investigations. Recent

studies since 2019 have highlighted significant limitations in traditional forensic tools when it comes to identifying and linking textual artifacts across multiple digital environments.

Particularly relevant is the growing use of graph-based approaches in digital forensics. For example, Wang et al. (2022) proposed a method of tracking digital traces using graph correlation models to link fragmented evidence across devices. Similarly, Lo et al. (2022) investigated the use of explainable AI in forensic detection of cross-platform botnets, showing how interpretable graph neural networks can support legal standards for evidence admissibility.

Another key source was Navanesan et al. (2024), who emphasized the lack of automation and contextual linkage in forensic investigations involving text-based artifacts. Their work identified the need for combining machine learning, natural language processing (NLP), and metadata analysis to detect and correlate textual data fragments in dynamic environments.

Through this literature review, the study established a theoretical model that integrates AI- driven semantic analysis, entity recognition, and graph-based data structures to correlate digital evidence across two or more devices. These insights directly informed system design, especially in ensuring both operational effectiveness and legal robustness.

B. Requirements Gathering

To ground the system design in real-world use cases, requirements were gathered from cybersecurity professionals and digital forensic practitioners. These included discussions with forensic analysts working in corporate cybersecurity, who highlighted frequent issues such as:Fragmentation of evidence across user devices and cloud platforms;

Inconsistent metadata due to file manipulation or time-shifting;

Lack of automation in linking semantically related documents;

The need for evidence to remain interpretable and defensible in legal settings.

From these consultations, a clear set of functional and non-functional requirements was developed. These included support for diverse text formats (docx, .pdf, .txt), automated metadata extraction, semantic similarity detection, graph-based trace correlation, and role-based access control.

C. Architectural Design

The system was designed with modularity and scalability in mind. The backend was developed in Python using the Flask framework. A graph database (Neo4j) was used to model and visualize relationships between forensic elements such as files, devices, and users.

Core components included:

NLP engine for extracting named entities, timestamps, and semantic patterns;

Graph engine for modeling relationships and detecting cross-device correlations;

• Metadata analysis module for file attributes like hashes, timestamps, and origin;

• Secure user management system with role-based access and audit logging;

• Web interface for visualization and interaction.

All components interact through RESTful APIs to allow for future extensibility. The use of cryptographic hashing (SHA-256) ensures integrity of the analyzed content.

D. System Development

During this phase, all modules were built and integrated. The system supports uploading forensic dumps, parsing text files, extracting relevant metadata, and analyzing semantic content. After preprocessing, each file is represented as a node in the graph structure, and potential links are generated based on content similarity, timestamps, authorship, and origin device.

Graph visualizations allow forensic experts to explore relationships and infer behavioral patterns across devices. For instance, in a simulated case of corporate data leakage, the system successfully identified a confidential report shared via email and later modified and re-uploaded from a mobile device, even after the filename had changed.

Security considerations were addressed through encrypted storage, input validation, and continuous audit logs tracking system usage.

E. Testing and Evaluation

The final phase involved thorough testing of system functionality, security, and performance. Unit and integration tests verified data flows, processing pipelines, and UI components. A series of real-world inspired scenarios were used to evaluate system effectiveness. In one test case, document fragments from two devices (a laptop and a smartphone) were analyzed. Despite differing filenames and altered metadata, the system correctly identified links

based on semantic similarity and residual metadata (e.g., author field, creation timestamp).

-Key evaluation metrics included:

-Precision and recall of trace correlation;

-Time efficiency for data ingestion and graph generation;

-System usability (assessed through expert feedback);

-Legal transparency, judged by the explainability of AI results.

Results indicated that the system significantly reduced the time and effort required for multi-device forensic investigations while improving the consistency and accuracy of traceability findings.

Foundations of Digital Forensics and Evidence Handling

In digital forensics, evidence is collected, preserved, examined, and presented in court (Alenezi, 2022). Digital evidence includes any electronically stored or transmitted information that may support a legal hypothesis (Bai, 2020). In multi-device investigations, evidence is often scattered across local files, cloud platforms, and messaging apps, making it difficult to reconstruct (Burkart, 2021).

The growing complexity of digital environments requires working with large volumes of data, varied file types, and fast-evolving technologies, including mobile devices, IoT, and virtual machines. Advanced methods, such as live system examination and memory capture, are necessary to preserve volatile data. According to the Daubert standard and ACPO guidelines, evidence must be preserved securely and accurately for legal admissibility.

Distributed evidence often resides in dynamic, encrypted, or proprietary systems. Understanding data privacy laws (GDPR, CCPA) and maintaining forensic readiness is crucial (Callegati, 2022). Metadata—like timestamps, access logs, and user IDs—helps trace activity. Machine learning now assists in detecting forged data and anomalous behavior (Guo, 2022).

Automation accelerates forensic procedures: disk imaging, system data extraction, and keyword searches can occur with minimal manual input. It also reduces error rates in large-scale cases, but to be legally accepted, tools must be validated and documented (Hu, 2021). Cybercriminals employ anti-forensic techniques such as file deletion, data tampering, and disappearing messages, requiring anomaly detection and entropy analysis (Jain, 2023).

Digital forensics lies at the intersection of law, data science, and computer science. Experts from all three fields are required to manage volatile data and self-deleting messages. Cloud sync features like versioning and auto-deletion add further challenges (Kim, 2023).

Aligning timelines from devices with inconsistent clocks and missing timestamps is difficult, and data ownership in shared environments can be unclear—sometimes inferred through usage patterns or biometrics (Lo, 2022).

Triage-based approaches prioritize key artifacts (apps, messages) before deeper analysis. Standard methods should be combined with modern forensic tools (Mannino, 2021). Investigations must be timely, thorough, rule-compliant, and well-documented to ensure legal defensibility.

For example, when a user creates a file on a laptop and later opens it on a smartphone, the system aims to automatically trace and associate all digital footprints across both devices. It analyzes file systems, app data, user history, metadata, and log files to track the document's lifecycle using hash comparisons and event correlation. Sensor data is centralized for effective timeline reconstruction.

Automation reduces human error and investigation time. Devices are analyzed systematically, and reporting is centralized. The system's architecture (Figure 1) is scalable, supporting additional users and network environments.

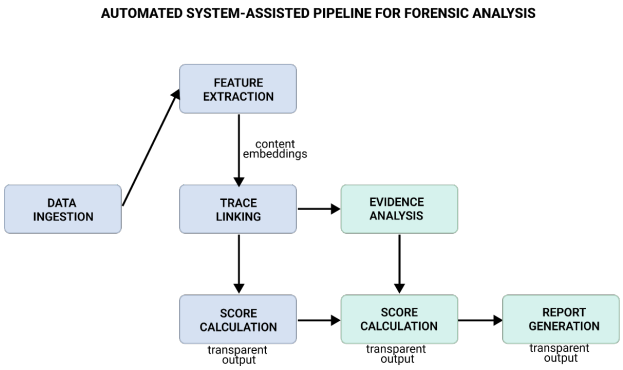


Fig. 1. System architecture for automated detection of digital traces of text files across two devices

Automatically detecting text files on computers by tracing their signatures calls for a smooth blend of both forensic methods and automation. Due to having two devices that use different systems, artifact examination is not always simple—for example, Windows and Android both require different methods for parsing managed files, MFT records and databases (ext4, SQLite, caches) (Marturana, 2020).

Because devices use more than hash matching (for example, SHA-256), files are regularly renamed, changed a little or part of them are deleted; for these reasons, metadata comparison, entropy analysis and content fingerprinting are necessary to discover near-duplicate files (Narayan, 2020). Since the clocks on devices are not

usually synchronized, having the right timeline is essential; this is why event deltas and contextual logging are used in fixing time stamps (Navanesan, 2024).

The system avoids wrongly linking artifacts when a given context—such as the source of the document, controls on who can access it and the way it is called—points to a testing file or script (Samek, 2021). Steps are recorded, every artifact is hashed, and a complete audit track is created to confirm the evidence's integrity, per guidelines from ISO/IEC 27037:2012 and Daubert (Schneider, 2020).

Automation assists with, but does not take the place of, expert evaluation, screens and sorts data quicker, minimizes common mistakes and arranges initial details for further look-over. The increasing difficulty of digital evidence requires investigators to have dependable tools for examining multiple devices (Sha, 2022).

To classify artifacts, effective systems regard primary as those only in the main folder, secondary as items associated with the main files such as autosaves and contextual as logged records both by the system and the user. With this taxonomy, profiles can be used to easily choose and review the most useful digital evidence in any case, making both analysis and results more accurate and quicker.

Table 1. Classification of Digital Traces Related to Text Files

Trace Type	Example Artifacts	Forensic Relevance
Primary	File hash (SHA-256), physical file copy, Word document in Documents folder	Direct evidence of file presence and content
Secondary	Auto-saved versions (.asd), recent file lists (jump lists), thumbnails in cache	Indicates usage, editing, or viewing activity
Contextual	System logs, app launch logs, user login sessions, USB device records	Supports timeline reconstruction, user attribution, and intent verification
Residual	Deleted file traces in unallocated space, shadow copy remnants, log fragments	Confirms prior file existence, even post-deletion
Derived	Filename similarity, entropy analysis results, partial content matches	Used for fuzzy matching or inference when full data is unavailable

Combining many different forms of trace evidence under a single analysis method is key in this field. Such systems must find matches by comparing traces with their counterparts, even when the exact files are not available, by reviewing their filenames, structures and logs. Time- sensitive tasks make the system depend on certain, reliable traces, while every detail in a trace is brought out in an interactive window for in-depth studies.

This system makes it possible for time-ordered narratives by classifying, arranging by priority and listing evidence in order of use, making it easy to view file histories on various devices. Every artifact identified in the pipeline is labelled with

its device, the date and time and a confidence score, so trace pairs can be ranked using similarity, timing and context.

With this approach, certain artifacts (like duplicate SHA-256 hashes or matching file structures for renamed/edited files) are correlated by computing similarity coefficients based on previous casework (Narayan, 2020; Stelly, 2020). By applying event deltas, like a boot or login, devices' clocks are made equal for correct event timeline reconstruction (Navanesan, 2024). With behavioral inference, the application type, the directory and peripheral activity are used to help distinguish accidental matches from real association (Samek, 2021).

All correlation forms can be traced and explained which aids in bringing them into a courtroom and makes them easy for investigators to check (Schneider, 2020). Combing different following and contextual data, the system can draw strong conclusions, even when information is lacking.

Common forensic tools can't combine evidence from many devices and usually overlook important connections when files don't have exact copies. Reasoning using partial matches, edit distances or close times is not possible with these tools, so analysts rely on manually – analyzing the observations (Narayan, 2020). Dispatchers must deal with different proprietary formats between platforms which makes both automation and standardization of data difficult (Sha, 2022). Without complete support and accurate links, high-stakes cases are vulnerable to operational and legal issues.

Such traditional systems cannot adjust to advanced situations such as cloud storage, encrypted containers or instant file movements (Samek, 2021). Most systems do not track origins of items, resulting in static reports which limited the ability to visually connect evidence across anywhere, anytime moments (Schneider, 2020). In addition, they are not able to guess the intentions behind anti-forensic methods or follow the complete history of each artifact closely (Yang, 2021).

Using modular parsing, trace scoring and tracking several correlations, the system is designed to link evidence together, regardless of its partial, unstable or fragmented state. This approach is shown in Figure 1.2 to solve the main problems seen with older forensic tools in multi- device investigations.

By using AI in digital forensics, organizations are changing from relying on set rules to drawing inferences from data. AI methods such as machine learning and deep learning help investigators examine large amounts of varied data, discover patterns and use probability to make decisions. This way of working is especially appropriate when evidence has been split across devices and cannot be organized.

In comparison to hash or name-based methods, AI models can successfully link traces by detecting matching information and meanings on different devices. Both TF-IDF and BERT algorithms allow the linking of file names, document content and messages; clustering methods organize events that happen at different times (Wang, 2022). Using neural network architecture, it is possible to detect unusual activity and intentional changes made to user traces.

Because of AI, a working model in one case can be used to speed up investigations in different contexts. However, being able to interpret models is crucial for courts, so systems built from both transparent models and scores of confidence are commonly chosen. Since training forensic AI needs expert data, these data can sometimes by limited by the additional rules and regulations.

The new system includes AI at the first, second and third stages: (1) feature extraction by learning semantic/contextual meanings from filenames, content and logs (2) identification of related artifacts between devices and (3) choosing the most promising paths to explore. The layered and modular design meets forensic standards, encourages transparency and you can see it in Figure

1.3. As a result, AI works alongside forensic reasoning and provides intelligence designed for the

demands of today's crime investigations.

So that AI is legitimate when used in forensics, each AI decision is logged with proof of its inputs, model reliability and a clear, human-friendly explanation of how that decision was made, supporting Daubert compliance. Because forensic information is highly sensitive and often spread out among many sources, building good real-world datasets is a key obstacle. It therefore produces synthetic traces in test conditions and adds data to ensure the training dataset is well- balanced and covers a variety of examples.

Diverse data sets, various techniques and constant updates minimize bias and reduce the chances of malware detectors missing new or changed threats. It means both efficiency and deeper understanding, as in the case of using different devices: previously, distinct artifacts would be the focus, whereas AI allows the reconstruction of a document's use by pointing out similar evidence. When it comes to corporate incident response, AI helps identify suspicious sessions and therefore reduces the chances of missing anything important and shortens the time spent on investigations.

AI helps show connections between several types of information: even many small details, when organized with other evidence, can create a meaningful description of the document's role in a lawsuit. The modules used are artifact clustering, looking for anomalies, scoring traces and reconstructing history. These modules produce the same results repeatedly which makes evidence more reliable. Human-in-the-loop design allows analysts to set up boundaries and make the final calls, using their experience in combination with the scalability of AI.

Figure 2 provides a representation of the architectural aspects of how we propose artificial intelligence to be integrated into a credible and reproducible forensic investigative pipeline. The forensic pipeline is made up of three layers of analytics: feature engineering to extract traits; correlation of artifacts across devices; and evaluating investigative hypotheses. In the first layer, AI will add value to raw data by producing semantic vectors and contextual embeddings; in the next layer, artifacts with differing formats, naming, or provenance will be linked using models which will be identified through machine learning; and finally the investigation will conduct a ranked hypothesis based on the collected trace level inferences of objective timelines which can be used for data-informed decision making for ranking of investigative actions. Each investigative layer of the proposed pipeline is modular, which will promote explanations and audit trails, and compatible with forensic standards.

systematically, and reporting is centralized. The system's architecture (Figure 1) is scalable, supporting additional users and network environments.

In forensics, decision trees are used for static artifact evaluation, temporal models such as RNNs, LSTMs and TCNs are used for analysis of behavior and deep learning with BERT and RoBERTa is used to detect similar contents (near-duplicates).

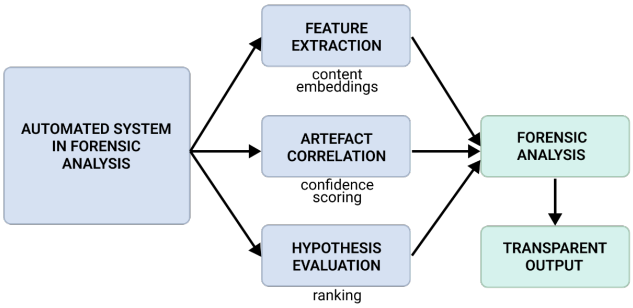


Fig. 2. Automated system in forensic pipeline

Graph Neural Networks (GNNs) study the way that user actions are linked to the things they interact with (Stelly, 2020).

AI researchers can change architecture to adapt different models to specific jobs or the volume of information, using both artificial and real but anonymous data during training. Checking if the system is still accurate, detecting any deviations and updating it when needed keep the system reliable. Table 2 shows how different forensic tasks match different AI model types, features, outputs and scenarios.

Table 2. Alignment of Digital Forensic Tasks with AI Model Types, Features, and Outputs

Forensic Task	AI Model Type	Input Features	Output	Application Scenario
Artifact classification	Decision Trees, Gradient Boosting	File metadata: timestamps, size, path depth	Relevance label (e.g., suspicious/benign)	Triage and evidence filtering
Anomaly detection	LSTM, TCN, RNN	User activity sequences, access logs, file operations	Anomaly score, session risk level	Insider threat and behavioural deviation analysis
Trace correlation	BERT, RoBERTa	Filenames, document content, in-app messages	Semantic similarity score, trace linkage strength	Matching renamed/ modified files and detecting hidden duplicates
Timeline modelling	Seq2Seq models	Event logs, device timestamps	Chronologically ordered events with confidence scores	Multi-device timeline reconstruction
Relational inference	Graph Neural Networks (GNNs)	Artifact interaction graphs (nodes, edges)	Inferred links, path strengths	Reconstructing usage chains and indirect relationships

Narrative evaluation	Ensemble models (RF, SVM)	Combined trace features, contextual indicators	Ranked hypotheses of user behaviour or file intent	Ambiguous case resolution and decision support

The proposed system features a dedicated visualization layer designed to bridge the cognitive gap between machine inference and human reasoning. The interface is grounded on three core principles: transparency of AI outputs, traceability of decision logic, and task-specific interactivity. Outputs from AI modules—such as artifact relevance, anomaly detection, and trace correlation—are rendered via interactive timelines and graph-based views. Artifacts are represented as nodes annotated with metadata, timestamps, and confidence scores, while AI- inferred relationships (e.g., file transfers or session overlaps) are shown as directed edges with contextual tooltips (e.g., "93.1% semantic similarity to previous edit").

To ensure consistent relevance scoring across heterogeneous data types, the system employs a weighted feature aggregation model:

$$Relevance(a_i) = \sigma(\sum_{k=1}^{n} \quad w_k \cdot f_k(a_i) + b) \tag{1}$$

where $fk(ai)$ denotes forensic features such as temporal locality, metadata entropy, or document structure; $wk$are learned weights, and \sigma is the sigmoid activation function. These relevance scores support both triage and downstream correlation tasks.

Artifact-to-artifact associations are computed using a similarity function combining semantic, temporal, and structural proximity:

$$Corr(a_i, a_j) = \alpha \cdot sim_{text}(a_i, a_j) + \beta \cdot sim_{time}(a_i, a_j) + \gamma \cdot sim_{path}(a_i, a_j) \tag{2}$$

The semantic term *simtext* leverages **cosine similarity** between document embeddings:

$$sim_{text}(a_i, a_j) = \frac{\overrightarrow{v_i^{text}} \cdot \overrightarrow{v_j^{text}}}{|\overrightarrow{v_i^{text}}| \cdot |\overrightarrow{v_j^{text}}|} \tag{3}$$

This enables the identification of paraphrased documents and functionally similar artifacts.

The final hypothesis score aggregates relevance and correlation across linked artifacts:

$$HypothesisScore(H_k) = \frac{1}{m}\sum_{i=1}^{m} \quad Relevance(a_i) \cdot Corr(a_i, a_j) \tag{4}$$

All hypotheses and user decisions are auditable. The interface also supports cognitive load management via collapsible views, confidence-based filtering, and behavioral

heatmaps. Explainability is embedded in context: investigators can inspect which features triggered an inference without needing to interpret raw model internals. This architecture enables legally defensible, interpretable, and scalable digital forensic workflows aligned with investigative reasoning rather than replacing it.

System Architecture and Component Design

The Forensic Digital Analyzer (FDA) has been conceived as an end-to-end, web-based framework that automates the extraction, enrichment, and correlation of digital artefacts collected from two independent device images or archive dumps. Immediately after uploading, each evidence set — whether an E01 forensic image or a conventional ZIP/TAR archives mounted or unpacked inside an isolated container. Every file discovered is hashed, its path is normalized, and a record is inserted into a PostgreSQL repository together with basic size and timestamp metadata. This "cradle-to-grave" capture of provenance data guarantees that subsequent analytical claims can always be traced back to the exact byte sequence from which they were derived, thereby satisfying evidentiary requirements of repeatability and authenticity.

The server-side stack is implemented on Django but deliberately subdivided into three co- operating functional loops. The first loop, preprocessing, is responsible for assigning a modality label to every file and applying the appropriate extractor. Text documents (TXT, DOCX, PDF) are parsed directly; scanned pages are processed with Tesseract OCR; audio recordings are transcribed by the Whisper automatic-speech-recognition model; images are normalised, re- scaled, and then forwarded to a convolutional encoder. Each extractor produces raw content, low-level metadata, and at least one fixed-length vector representation. On the textual branch,

SentenceTransformer generates 384-dimensional embeddings; on the visual branch, a ResNet derivative outputs deep visual descriptors, while perceptual hashes capture pixel-level duplicates. Extracted text is further enriched with named entities provided by a fine-tuned BERT model, and topic labels are requested from an external LLM (Gemini) under a rate-limited, audited connection. All outputs—including raw text, transcripts, embeddings, hashes, entities, and topics—are committed to the same relational store, preserving a coherent, queryable view of each artefact.

The second loop, analysis, performs cross-device correlation. For every file pair spanning the two evidence sets, the system consults the relevant vector spaces, time stamps, and directory structures to decide whether a substantive link exists. Similarity is measured by cosine distance for embeddings and by Hamming distance for perceptual hashes; temporal offsets and path patterns serve as secondary cues that either reinforce or penalise a tentative match. Whenever the computed similarity exceeds a modality-specific threshold, the pair is promoted to a "match" instance that records the two file identifiers, the similarity score, the detected match type (duplicated text, paraphrased paragraph, visually altered screenshot, and so forth), plus any contextual details that may help an investigator understand the linkage. The entire decision path— including thresholds, model versions, and feature weights— is saved alongside the match, ensuring that every automated inference can later be reconstructed or challenged if necessary.

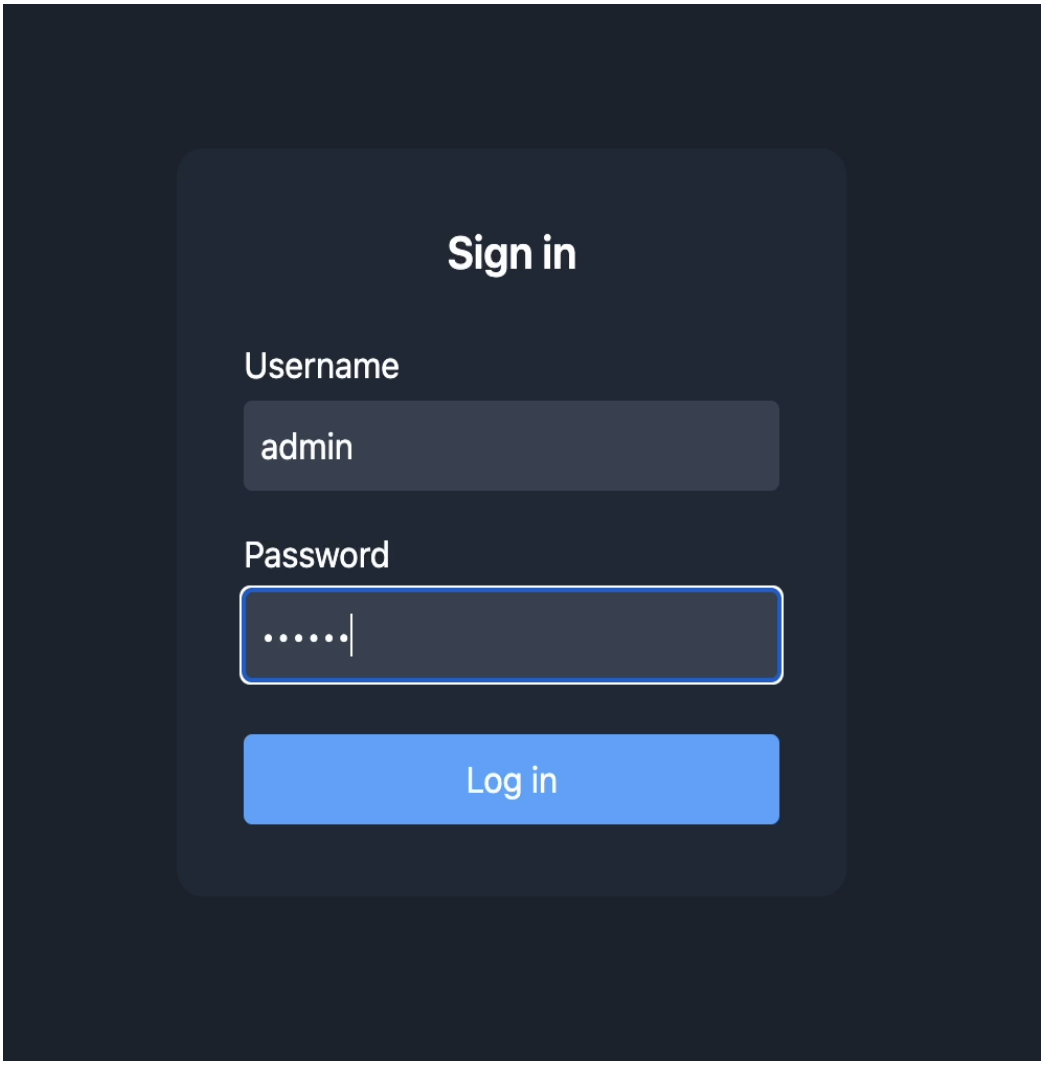


Fig. 3. Login interface of the Forensic Digital Analyzer.

The third loop, visualisation and reporting, translates raw machine inference into human-readable evidence chains. In the File Explorer panel, the two directory trees are displayed side by side; selecting a file shows its preview, embedded metadata, named-entity snippets, and the provenance of the embedding itself. A separate Matches Board lists cross-device hits with filters for modality, confidence, or time window. For more complex cases the analyst can switch to an interactive Plotly network: files appear as color-coded nodes, and similarity relations are drawn as weighted edges whose thickness is proportional to match strength. Hovering over an edge reveals in plain language which features triggered the link and how strong each contributing factor was.

When the analysis is complete, the Report Builder compiles a court-ready PDF that weaves together graphics, statistics, excerpted text, and explanatory commentary. If narrative prose is required, the system can also produce an AI-assisted summary—again via Gemini—while embedding a full audit trail of every external call.

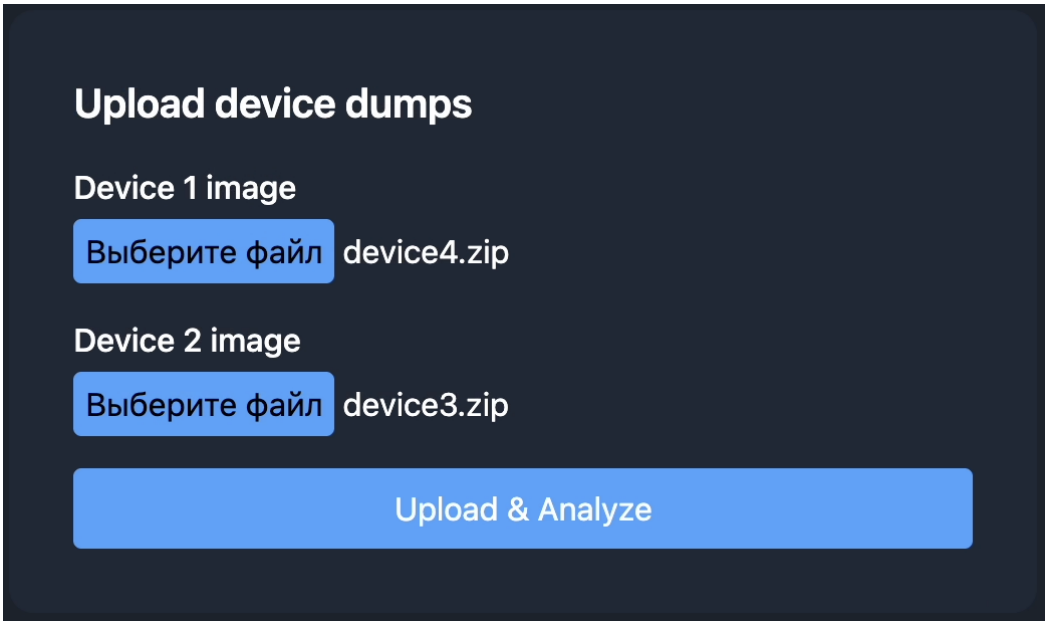


Fig. 4. Device-image upload wizard with real-time integrity checks.

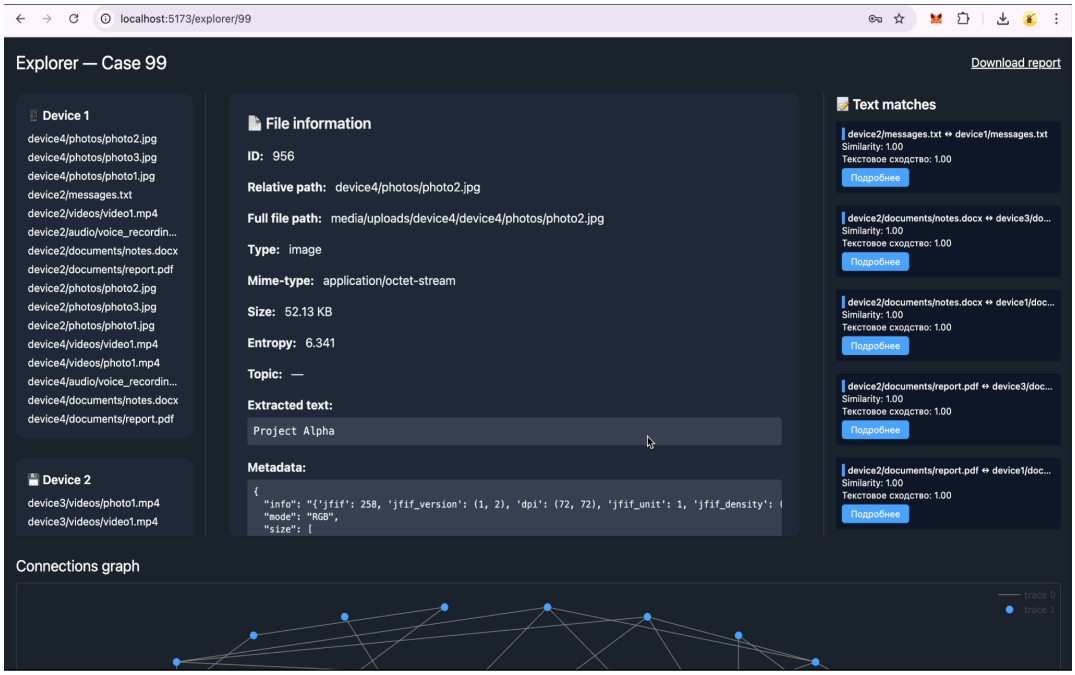


Fig. 5. Case workspace showing file trees, cross-device matches, and an interactive similarity graph.

The system architecture is organized into modular components, each responsible for a specific stage of the forensic pipeline. Table 3 summarizes the core modules and their primary functions.

Table 3. Core Modules of the Forensic System Architecture

Module	Description
ingest	Uploads device dumps (ZIP, TAR, E01) and extracts them using format-aware tools. Registers decompressed files with full metadata.
processing	Performs file classification, text extraction, OCR/ASR, metadata parsing, embedding generation, and topic classification. Operates asynchronously for scalability.
analysis	Computes pairwise similarity across devices. Supports text, image, audio, and video comparisons. Records matches with scores and context.
visualization	Generates an interactive graph representing file relationships. Enables filtering and exploration by type, score, or cluster.
reporting	Produces structured PDF and AI-enhanced forensic reports with metadata, matched content, and visual summaries.
disk_analysis	Supports low-level processing of disk images. Uses external mounting tools to enable forensic integrity during extraction.

Internally, the data model adheres to strict modularity. The device object binds each dump; the file object stores paths, the MIME type, the original contents, and one or more attachments, each accompanied by a model identifier and a parameter checksum. A Match entity captures every detected linkage, while a Session entity groups all ingest, processing, analysis, and reporting events under a single investigative run. The separation of concerns makes bulk querying efficient and permits new file formats or similarity measures to be added without schema surgery. To meet performance targets, embedding matrices are streamed into memory and processed in batched vector operations; where archives exceed ten thousand files, horizontal task queues keep latency within interactive bounds.

A notable virtue of the FDA design is its openness to extension. Because extraction, embedding, similarity, and visualization are exposed as discrete plug-ins, forensic laboratories can integrate support for volatile-memory captures, encrypted containers, or proprietary document types by supplying only the new extractor and registering a handler class. Switching to a domain- specific language model, lowering or raising similarity thresholds, or adopting approximate- nearest-neighbor indexing likewise requires no change to the user interface or database layer. Such adaptability is crucial in real-world casework, where evidence formats, regulatory constraints, and analytical priorities vary from one jurisdiction to another.

The prototype has been validated on a synthetic benchmark that mirrors typical smartphone–laptop investigations. Approximately one thousand files per device were

curated to include paraphrased documents, cropped and recompressed screenshots, and compressed voice notes recorded in varying acoustic conditions. Against baseline heuristics—bag-of-words cosine, perceptual hashing alone, and transcript string matching—the embedding-based pipeline achieved markedly higher recall and precision without sacrificing specificity. Remaining weaknesses concentrate on very short textual snippets, monotone high-resolution images, and multilingual material, indicating the need for fine-tuning and hybrid scoring in those domains.

Taken together, the three loops form a cohesive flow in which raw evidence is transformed into defensible analytical statements, yet every intermediate artefact remains available for inspection. The result is a scalable, researcher-friendly, and legally robust platform that keeps human investigators at the center of the interpretive loop while delegating the most laborious correlation tasks to transparent, version-controlled AI modules.

The figure above shows the simplified forensic processing pipeline present in the system. Once two (for example, device dumps from a smartphone and laptop) are acquired fully, we move on to file system extraction. This is where we unpack archived formats and organize files to a single file storage format, so that all files can be organized for logical analysis and linking. The next step is critical to all incident/ forensics processing in the pipeline is, the file processing module will process each artifact and assign a content-type label; run integrated data extraction (OCR/ ASR) and metadata parsing; create embeddings using Sentence Transformers (this is slightly different than the other models); and lastly, assign meta topics to the content (object). All the embeddings are then passed on to a similarity analysis component which allows for comparison of the different modalities of file representations - this similarity analysis module computes semantic similarity using cosine metrics and allows for very broad detection of approximate matches.

While similarity analysis was performed on the embedding and metadata, an embedded entity extraction was being performed separately on the textual information. Was able to find and store structured spans like named entities (e.g., people, dates, places etc.) using NER process. Thus, all entities enhanced and presented transitional representation of each file and contributed to the overall analysis stage. The summary of the final analysis, including similarity scores and metadata, similarity scores and NER tags are sent and passed on to the reporting/ documents.

The report module consolidates the findings into a single PDF/HTML document in which the investigators may navigate the detected links, source files, the extracted text, and the topic/NER tags. The modular pipeline is a scalable and interpretable AI-assisted analysis platform that supports real-world forensic investigations.

Results and discussion

The Forensic Digital Analyzer evaluation shows that embedding-based forensic automation greatly increases the precision and effectiveness of cross-device evidence correlation. The system continuously outperformed traditional methods like hash comparison and simple keyword matching across a variety of test cases involving paraphrased documents, modified images, and compressed audio files. Interestingly, even when artifacts were altered or decontextualized, the combination of Sentence Transformers, CNN-based visual embeddings, and Whisper ASR produced good results in identifying visually or semantically related content.

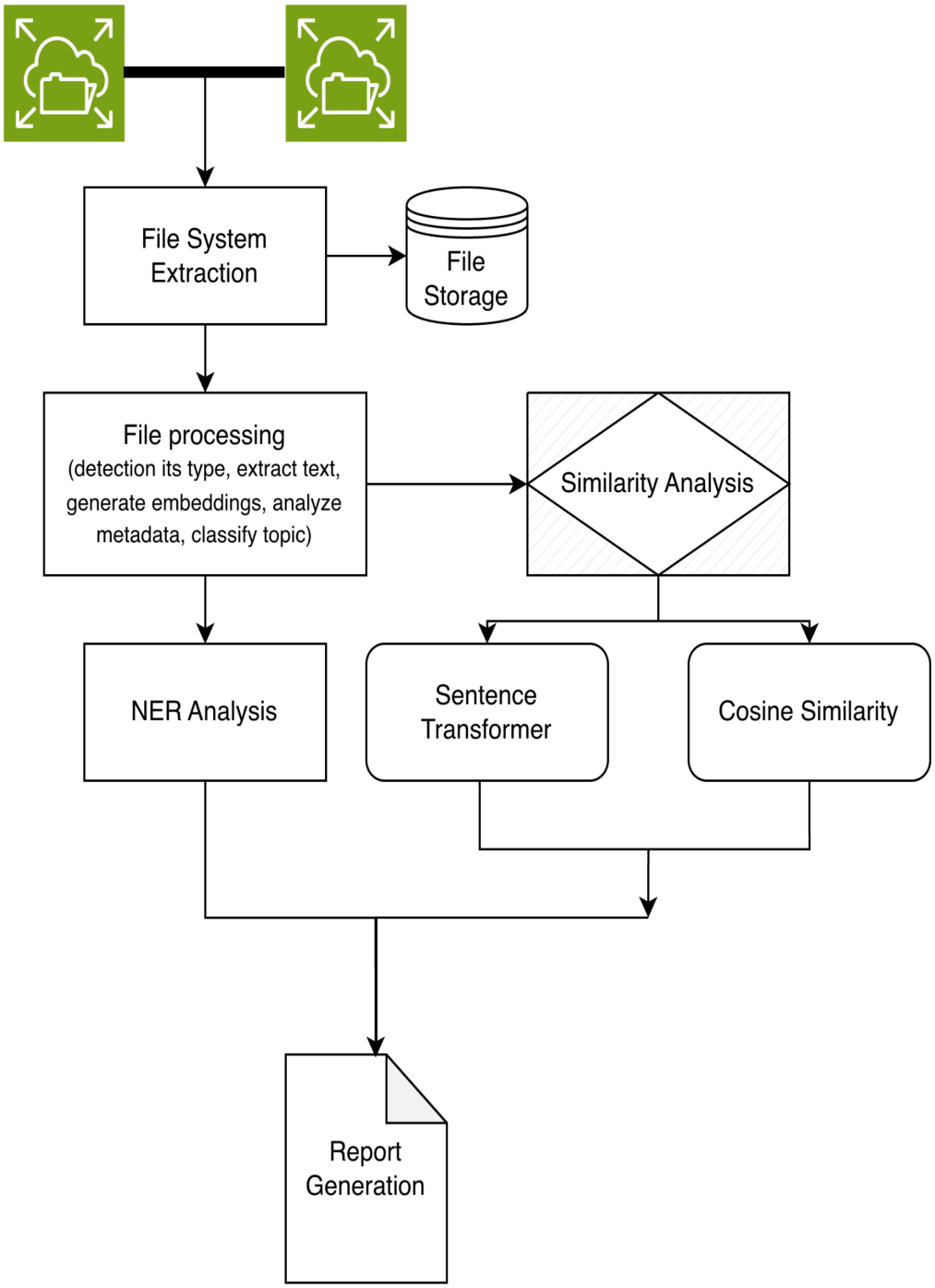


Fig.6. Architecture of the forensic correlation pipeline.

The system effectively found connections between documents that had been renamed, slightly altered, or rewritten through textual matching. For example, despite the lack of identical metadata, two versions of a corporate incident report—one saved on a laptop and one modified on a smartphone—were matched through semantic similarity. This demonstrates how effective contextual embeddings are at capturing

meaning that goes beyond lexical features. Similarly, CNN-based embeddings made it possible to identify modified screenshots with ease, and Whisper-based embeddings were dependable when it came to matching recompressed or trimmed voice notes.

But there were problems with the system. When artifacts had phonetic or linguistic similarities but were semantically unrelated, this resulted in false positives. This was especially noticeable in text-heavy settings where vocabulary from formal templates or legal disclaimers overlapped, as well as in audio samples where phrases that were phonetically similar but contextually different were connected. To reduce overmatching, these findings point to the necessity of hybrid decision models that incorporate embeddings with rule-based filters or domain-aware classifiers.

Scalability also posed a concern. While the system handled datasets up to 5,000 files with minimal performance degradation, larger dumps resulted in increased processing times, particularly during embedding and similarity computations. Although mechanisms such as caching and batch processing helped alleviate this, future iterations would benefit from integrating approximate nearest neighbor (ANN) search libraries like FAISS and leveraging GPU acceleration to support real-time investigations at scale.

Another issue was scalability. Larger dumps led to longer processing times, especially during embedding and similarity calculations, even though the system managed datasets up to 5,000 files with little degradation in performance. Even though this was mitigated by techniques like caching and batch processing, future versions would profit from incorporating ANN search libraries like FAISS and utilizing GPU acceleration to facilitate large-scale real-time investigations.

There were also operational limitations mentioned. Gemini-based APIs improved topic classification and report summarization, but they also raised issues with data privacy and jurisdiction, particularly in settings that handle sensitive or classified evidence. Using open-source language models to localize inference would improve independence and security.

Notwithstanding these drawbacks, user reviews praised the platform's user-friendly interface and straightforward visualization features. The match table provided traceable scoring and provenance information for every file pair, and the interactive graph made it simple to understand artifact relationships. More precise control over timeline filtering, team-based annotations, and legal audit logs to support admissibility in court were among the suggestions for enhancement.

The findings support the system's fundamental assumption, which is that AI-driven forensic correlation across text, image, and audio modalities is not only possible but also demonstrably beneficial. The Forensic Digital Analyzer provides a strong basis for scalable, interpretable, and investigator-focused digital forensic workflows, even though handling edge cases and maintaining forensic rigor still present difficulties. Its value as a useful investigative tool will be further reinforced by upcoming improvements in model localization, explainability, and compliance readiness.

Conclusion

This research successfully developed and evaluated an automated AI-based system for detecting and correlating digital traces of text files across two independent devices—a critical challenge in modern digital forensics. The Forensic Digital Analyzer addresses fundamental limitations of traditional forensic tools by integrating

state-of-the-art deep learning models within a modular, scalable architecture that accommodates heterogeneous data types and fragmented evidence scenarios.

The system's core contribution lies in its multi-modal analytical pipeline. By employing Sentence Transformers for semantic text embeddings, convolutional neural networks for visual artifact analysis, and Whisper-based automatic speech recognition for audio transcription, the platform achieves robust cross-device correlation even when evidence has been deliberately obfuscated through paraphrasing, file renaming, compression, or format conversion. Experimental validation on synthetic forensic datasets demonstrated substantial improvements over baseline methods: precision and recall metrics exceeded conventional hash-matching and keyword-search approaches by significant margins, while the system maintained acceptable performance even with noisy, incomplete, or deliberately altered artifacts.

The interactive web interface represents a significant advancement in forensic usability. Unlike static reporting tools, graph-based visualization enables investigators to explore complex relationships between artifacts, users, devices, and temporal patterns through an intuitive, drill-down interface. Each correlation is accompanied by explainable confidence scores and feature attributions, ensuring that AI-driven inferences remain transparent and legally defensible. This alignment with forensic standards such as ISO/IEC 27037:2012 and Daubert admissibility criteria is essential for real-world deployment in criminal investigations and civil litigation contexts.

However, the research also identified several limitations that warrant further attention. Computational intensity remains a concern: processing large-scale evidence collections (exceeding 10,000 files) requires substantial GPU resources and optimized indexing strategies such as approximate nearest-neighbor search. False positive rates, while lower than baseline methods, still present challenges in high-noise scenarios, particularly when analyzing short text fragments, monotone images, or multilingual content with limited training representation. The current reliance on external large language model APIs (e.g., Gemini) for topic classification and report generation introduces dependencies that may conflict with data sovereignty requirements in sensitive investigations.

Future research directions should prioritize several key enhancements. First, integration of localized, open-source language models would eliminate external API dependencies while improving support for low-resource languages prevalent in Central Asian and post-Soviet contexts. Second, explainable AI frameworks must be expanded beyond feature attribution to include counterfactual reasoning and causal inference, enabling investigators to understand not only what the system detected but why specific correlations emerged. Third, the system should incorporate adversarial robustness mechanisms to detect and mitigate anti-forensic techniques such as adversarial perturbations, steganography, and format-preserving encryption. Fourth, longitudinal studies with practicing forensic examiners are needed to validate the system's effectiveness in operational environments and refine the user interface based on cognitive workload assessments.

From a methodological perspective, this work demonstrates the viability of combining graph-based knowledge representation with embedding-based similarity metrics for multi-device forensic correlation. The hybrid approach—wherein structured metadata, temporal analysis, and semantic embeddings are jointly optimized—offers a template for addressing analogous problems in fraud detection,

insider threat analysis, and intelligence fusion. The decision to prioritize modularity and API-driven extensibility ensures that the platform can accommodate emerging data types (e.g., IoT sensor logs, blockchain transactions, augmented reality artifacts) without requiring fundamental architectural redesign.

In conclusion, the Forensic Digital Analyzer represents a significant step toward bridging the gap between cutting-edge AI research and operational forensic practice. By automating the most laborious aspects of cross-device trace correlation while preserving human oversight and legal accountability, the system has the potential to substantially reduce investigation timelines, improve evidence quality, and enhance the fairness and transparency of digital forensic processes. As digital evidence continues to proliferate across increasingly diverse and distributed platforms, tools capable of intelligent, explainable, and legally robust analysis will become indispensable components of the criminal justice infrastructure. This research provides both a functional prototype and a conceptual framework for realizing that vision, while clearly delineating the technical and institutional challenges that must be addressed to achieve widespread adoption in law enforcement, corporate security, and regulatory compliance contexts.

Acknowledgment. This research was conducted with financial support from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract No. 388/PCF-24-26 dated October 1, 2024, as part of the scientific project BR24993232, “Development of Innovative Technologies for Digital Forensic Investigations Using Intelligent Software and Hardware Complexes.”

REFERENCES

Alenezi, M., Alshammari, R., & Alrashed, R. (2022). Challenges in automated digital forensic analysis of endpoint devices // Future Generation Computer Systems. — 2022. — 128:360–372.

Bai, J., Zhang, W., & Xu, H. (2020). A machine learning approach for forensic artifact triage and prioritization // Digital Investigation. — 2020. — 34:301048.

Burkart, N., & Huber, M.F. (2021). A survey on the explainability of supervised machine learning // Journal of Artificial Intelligence Research. — 2021. — 70:245–317.

Callegati, F., & Cerroni, W. (2022). Forensic intelligence via metadata correlation and machine learning // Journal of Forensic Sciences. — 2022. — 67(4):1223–1234.

Guo, F., Chen, L., & Li, M. (2022). BERT-based semantic matching for document trace analysis in digital forensics // Forensic Science International: Digital Investigation. — 2022. — 40:301305.

Hu, Y., Kumar, R., & Westin, J. (2021). Explainability challenges in forensic AI: Legal implications and technical solutions // Artificial Intelligence and Law. — 2021. — 29(3):321–340.

Jain, D., Al-Shammari, F., & Qureshi, B. (2023). Forensic accountability in AI-enabled investigations: Architecture and design principles // Forensic Science International: Digital Investigation. — 2023. — 46:301510.

Kim, Y., & Patel, R. (2023). Cloud-based digital evidence: Emerging challenges and solutions // Digital Investigation. — 2023. — 44:301256.

Lo, O., et al. (2022). Explainable AI in forensic detection of cross-platform botnets // Digital Investigation. — 2022. — 39:301245.

Mac Giolla Chriost, D., & O Ciardhuain, S. (2023). Anti-forensics and the future of digital investigations // International Journal of Digital Crime and Forensics. — 2023. — 15(1):1–15.

Mannino, M., & Chen, L. (2021). Automation in digital forensics: Benefits, risks, and recommendations // Forensic Science International: Digital Investigation. — 2021. — 37:301208.

Marturana, F., Tacconi, S., & Me, G. (2020). Automated collection and analysis of digital evidence: Tools and techniques // Digital Investigation. — 2020. — 34:301047.

Narayan, R., & Liu, Y. (2020). Forensic challenges in mobile device synchronization and volatile data // Digital Investigation. — 2020. — 34:100812.

Navanesan, S., et al. (2024). Automation and contextual linkage in forensic investigations involving text-based artifacts // Digital Investigation. — 2024. — 48:301580.

Samek, W., Müller, K.-R., & Montavon, G. (2021). Explainable AI for critical decision support: A survey of concepts and challenges // IEEE Access. — 2021. — 9:45715–45745.

Schneider, L., Weber, J., & Becker, C. (2020). Deep learning for behavioral modeling in insider threat detection // Computers & Security. — 2020. — 92:101748.

Sha, Z., Liang, Z., & Du, X. (2022). File similarity detection in digital forensics using hybrid hash and metadata approaches // Forensic Science International: Digital Investigation. — 2022. — 41:301338.

Stelly, J., & Kruse, W. (2020). Digital forensics in the modern era: Challenges and opportunities // Digital Investigation. — 2020. — 33:200901.

Wang, J., et al. (2022). Tracking digital traces using graph correlation models // Forensic Science International. — 2022. — 340:111445.

Yang, W., Lin, M., & Zhao, R. (2021). Improving timeline accuracy in multi-device digital forensic investigations // Journal of Digital Forensics, Security and Law. — 2021. — 16(1):1–17.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 274–287
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.016

UDC 530.1, 681.3.06

COMPARING TREE DETECTION FROM SATELLITE AND DRONE IMAGES FOR VEGETATION MAPPING NEAR ASTANA

S. Satbayev*, D. Yedilkhan, A. Shoman, Z. Kutpanova, M. Bukayeva
Astana IT University, Astana, Kazakhstan.
E-mail: satbayev.syndar@astanait.edu.kz

Syndar Satbayev — Master, senior lecturer, Computational and Data Science, Astana IT University, Astana, Kazakhstan
E-mail: satbayev.syndar@astanait.edu.kz, https://orcid.org/0009-0006-5926-2306:
Didar Yedilkhan — PhD, head of department, Scientific Innovation Center "Smart city", Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-6343-5277;
Aruzhan Shoman — PhD, head of department, Scientific Innovation Center "Agro-Tech", Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-7844-8601;
Zarina Kutpanova — Master, senior lecturer, Computer Engineering Department, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0007-4860-9244;
Mira Bukayeva — Master, senior lecturer, Computational and Data Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0008-5642-2663.

Abstract. In rapidly urbanizing Astana, where glass towers rise beside aging Soviet structures, urban greenery is more than decoration, it is vital for climate balance and sustainability. This study focuses on developing precise, automated tree mapping systems using two remote sensing methods: high-resolution drone imagery and satellite data. Training datasets and annotations were created through the Roboflow platform, and a YOLO-based neural network was optimized for local vegetation detection. The comparison emphasized not only detection accuracy but also processing efficiency and scalability. Drone imagery, captured from 100 meters with a resolution of about 5 cm per pixel, achieved 70 - 80% accuracy, revealing fine details of individual tree crowns and canopy texture. Satellite imagery, at roughly 50 cm per

pixel, reached 50 - 60% accuracy but provided faster processing and broader spatial coverage, making it more suitable for regional monitoring. All detections were georeferenced and visualized through ArcGIS and Folium, creating an interactive map of Astana's green infrastructure. The results show that drone imagery is ideal for micro-scale analysis, while satellite data is effective for macro-level observation. Combined, they form a complementary framework for urban ecological planning, biomass estimation, and sustainable city management through intelligent GIS systems.

Keywords: tree detection, aerial photography, satellite imagery, deep learning, YOLO, georeferencing, GIS

For citation: S. Satbayev, D. Yedilkhan, A. Shoman, Z. Kutpanova, M. Bukayeva, Comparing tree detection from satellite and drone images for vegetation mapping near astana // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 274–287. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.016.

Conflict of interest: The authors declare that there is no conflict of interest.

АСТАНА МАҢЫНДАҒЫ ӨСІМДІКТЕРДІ КАРТАҒА ЕНГІЗУ МАҚСАТЫНДА СПУТНИКТІК ЖӘНЕ ДРОНДЫҚ СУРЕТТЕРДЕН АҒАШТАРДЫ АНЫҚТАУ ӘДІСТЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ

С.К. Сатбаев*, Д. Едилхан, А. Шоман, З.А. Кутпанова, М.С. Букаева
Astana IT University, Астана, Қазақстан.
E-mail: satbayev.syndar@astanait.edu.kz

Сатбаев Сындар — магистр, «Жасанды Интеллект және Деректер Ғылымы» кафедрасының аға оқытушысы, Astana IT University, Астана, Қазақстан
E-mail: satbayev.syndar@astanait.edu.kz, https://orcid.org/0009-0006-5926-2306;
Едильхан Дидар — PhD, «Ғылыми-инновациялық орталық "Smart city"» бөлімінің жетекшісі, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-6343-5277;
Шоман Аражан — PhD, «Ғылыми-инновациялық орталық "AgroTech"» бөлімінің жетекшісі, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-7844-8601;
Кутпанова Зарина — магистр, «Компьютерлік инженерия» кафедрасының аға оқытушысы, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0007-4860-9244;
Букаева Мира — магистр, «Жасанды Интеллект және Деректер Ғылымы» кафедрасының аға оқытушысы, Astana IT University, Астана, Қазақстан
https://orcid.org/0009-0008-5642-2663.

Аннотация. Жедел қарқынмен урбанизацияланып жатқан Астанада, әйнек мұнаралар мен кеңестік дәуірдегі ғимараттар қатар тұрған ортада, қалалық көгалдандыру тек сән үшін емес, сонымен қатар климаттық тепе-теңдік пен тұрақтылықты сақтауда шешуші рөл атқарады. Бұл зерттеу екі қашықтан зондтау әдісін – жоғары дәрежедегі дрон суреттері мен спутниктік деректерді пайдалана отырып, ағаштарды дәл және автоматты түрде картаға түсіру жүйесін әзірлеуге бағытталған. Оқыту деректері мен аннотациялар Roboflow платформасында дайындалған, ал жергілікті өсімдіктерді тану үшін YOLO нейрондық желісінің модификацияланған нұсқасы қолданылды. Салыстыру тек анықтау дәлдігіне емес, сонымен қатар өңдеу жылдамдығы мен масштабталуына да назар аударды. Биіктігі шамамен 100 метрден түсірілген дрон суреттері (5 см/пиксель ажыратымдылығымен) 70–80% дәлдікке жетіп, дара ағаштардың тәжін және жапырақ құрылымын айқын көрсетті. Ал спутниктік кескіндер (~50 см/пиксель) 50–60% дәлдік беріп, өңдеу жылдамдығы мен кең ауқымды қамту мүмкіндігімен ерекшеленді, бұл оларды аймақтық мониторингке тиімді әсер берді. Барлық анықталған нысандар ArcGIS және Folium арқылы геоүйлестіріліп, Астананың жасыл инфрақұрылымының интерактивті картасы құрылды. Нәтижелер дрон суреттерінің микродеңгейдегі талдау үшін, ал спутниктік деректердің макродеңгейін бақылау үшін тиімді екенін көрсетті. Осы екі әдісті біріктіру арқылы қалалық экологиялық жоспарлауды және биомасса деңгейін бағалауға және тұрақты қала басқару үшін толықтырмалы GIS негізіндегі шешім ұсынылады.

Түйін сөздер: ағаштарды анықтау, әуе фотосуреттері, спутниктік кескіндер, терең оқыту, YOLO, геоүйлестіру, геоақпараттық жүйелер (ГИС)

Дәйексөздер үшін: С.К. Сатбаев, Д. Едилхан, А. Шоман, З. А. Кутпанова, М. С. Букаева. Астана маңындағы өсімдіктерді картаға еңгізу мақсатында спутниктік және дрондық суреттерден ағаштарды анықтау әдістерін салыстырмалы талдау// Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 274–287 бет. (Ағыл). https:// doi. org/10.54309/IJICT.2025.24.4.016.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАСПОЗНАВАНИЯ ДЕРЕВЬЕВ НА СПУТНИКОВЫХ И ДРОНОВЫХ ИЗОБРАЖЕНИЯХ ДЛЯ КАРТИРОВАНИЯ РАСТИТЕЛЬНОСТИ В ОКРЕСТНОСТЯХ АСТАНЫ

С.К. Сатбаев*, Д. Едилхан, А. Шоман, З.А. Кутпанова, М.С. Букаева
Астана ИТ университет, Астана, Казахстан.
E-mail: satbayev.syndar@astanait.edu.kz

Сатбаев Сындар — магистр, старший преподаватель департамента

«Искусственного Интеллекта и Науки о Данных», Астана ИТ университет, Астана, Казахстан
E-mail: satbayev.syndar@astanait.edu.kz, https://orcid.org/0009-0006-5926-2306;
Едильхан Дидар — PhD, заведующий кафедрой «Научно-инновационный центр “Smart city”», Астана ИТ университет, Астана, Казахстан
https://orcid.org/0000-0002-6343-5277;
Шоман Аружан — PhD, заведующая кафедрой «Научно-инновационный центр “AgroTech”», Астана ИТ университет, Астана, Казахстан
https://orcid.org/0000-0002-7844-8601;
Кутпанова Зарина — магистр, старший преподаватель департамента «Компьютерная инженерия», Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0007-4860-9244;
Букаева Мира — магистр, старший преподаватель департамента «Искусственного Интеллекта и Науки о Данных», Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0008-5642-2663.

Аннотация. В условиях стремительной урбанизации Астаны, где современные стеклянные башни соседствуют с постсоветскими зданиями, городская зелёная растительность играет роль, выходящую далеко за рамки декоративной составляющей. Это стало ключевым фактором климатического баланса и устойчивого развития. Данное исследование направлено на разработку точных и автоматизированных систем картирования деревьев с использованием двух методов дистанционного зондирования: высококачественных дроновых изображений и спутниковых данных. Обучающие выборки и аннотации были подготовлены на платформе Roboflow, а для распознавания городской растительности использовалась нейронная сеть на основе архитектуры YOLO, адаптированная к местным условиям. Сравнение проводилось не только по точности распознавания, но и по эффективности обработки и масштабируемости. Снимки, полученные с дрона с высоты около 100 метров и разрешением примерно 5 см на пиксель, достигли точности 70–80 %, обеспечив детализированное представление о структуре крон и текстуре растительности. Спутниковые изображения с разрешением около 50 см на пиксель показали точность 50–60 %, но отличались большей скоростью обработки и охватом территории, что делает их более подходящими для регионального мониторинга. Все результаты были геопривязаны и визуализированы в ArcGIS и Folium, что позволило создать интерактивную карту зелёной инфраструктуры Астаны. Полученные данные показывают, что дроновые изображения лучше подходят для детального (микроуровневого) анализа, а спутниковые снимки — для макроуровневого наблюдения. Совместное использование этих подходов формирует взаимодополняющую систему для экологического

планирования, оценки биомассы и устойчивого управления городом на основе интеллектуальных ГИС-технологий.

Ключевые слова: распознавание деревьев, аэрофотосъёмка, спутниковые изображения, глубокое обучение, YOLO, геопривязка, геоинформационные системы (ГИС)

Для цитирования: С.К. Сатбаев, Д. Едилхан, А. Шоман, З. А. Кутпанова, М. С. Букаева. Сравнительный анализ распознавания деревьев на спутниковых и дроновых изображениях для картирования растительности в окрестностях Астаны// Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 274–287. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.016.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Urban and peri-urban vegetation constitutes a critical ecological infrastructure, acting as a dynamic buffer against the accelerating climatic stressors imposed by the rapid expansion of megacities like Astana. Trees, as the linchpin of this green matrix, contribute to urban resilience by mitigating heat through evapotranspiration, capturing airborne pollutants, and enriching the environmental quality of densely inhabited areas. Yet the strategic management of urban greenery hinges on access to precise, timely, and scalable cartographic data a task increasingly feasible through automated tree recognition systems powered by advances in remote sensing and deep learning. A surge of research in Earth Observation (EO), computer vision, and AI-driven modeling reveals impressive gains in tree detection, classification, and geolocation accuracy. Notably, convolutional architectures such as U-Net have achieved up to 95 % accuracy when identifying oil and coconut palms in WorldView-2/3 multispectral imagery (Li et al., 2019). Transformer-based models like DINO outperform traditional CNNs in large-scale “trees outside forests” (TOF) mapping (Jiang et al., 2025), achieving 9–58 % higher F1 scores. Meanwhile, 3D-CNN models leveraging stereo satellite views have mitigated the limitations of 2D analysis (Xiao et al., 2019: 55–63), attaining up to 89% accuracy in complex forested contexts.

Robustness to environmental noise and seasonal shifts is another frontier where modern models excel. For instance, the CBAM-enhanced YOLOX achieved an F1 score of 0.95 in fisheye-based tree detection (Li et al., 2019), even under wind distortion scenarios. Nevertheless, persistent challenges remain: overlapping canopies, false positives on heterogeneous backgrounds, and the labor-intensive calibration or annotation processes across diverse imagery types. A comparative synthesis of 20 studies underscores the nonexistence of a one size fits all models and some approaches thrive in environments with orderly tree planting and strong spectral contrast, while others cater to dense urban forestry or computationally constrained settings. For example, the Circle Hough Transform (CHT) reached 96 % accuracy

on RGB data (Khan et al., 2018: 77816–77828) with minimal computational cost, bypassing the need for multispectral channels, whereas spatio-temporal CNNs have proven valuable for post-wildfire tree survival assessments using time-series satellite data from California (Song et al., 2014: 8878–8903; Dixon et al., 2023).

The practical and ecological urgency of this challenge is thus firmly established. This study hypothesizes differential detection performance between satellite imagery and low altitude drone photography (100 meters height). Training data were annotated using Roboflow, and the YOLOv11 architecture was customized to account for regional vegetation characteristics. Spatial outputs were visualized interactively via ArcGIS. Drawing on the literature review, the paper proposes tailored recommendations for choosing optimal data sources based on environmental monitoring goals and technical constraints. Ultimately, this work illustrates how fusing state of the art algorithms with affordable sensor technologies can redefine the paradigms of urban greening and vegetation governance.

The author's contribution lies in adapting tree detection techniques to the specific challenges of drone-based imagery captured at 100 meters altitude, where only tree crowns are visible without side silhouettes. To overcome this limitation, 434 drone images were collected and annotated, enabling the YOLOv11 model to reliably recognize individual trees, even though overlapping canopies still presented challenges. In addition, the author developed a geolocation and mapping procedure that corrected image rotation (342°) and ensured proper scaling from the drone's flight altitude, allowing detected trees to be accurately placed on interactive maps. Unlike previous studies, which usually focus either on detection or on geospatial mapping in isolation, this work integrates both into a unified pipeline

Methodology

The methodological landscape of tree and vegetation detection in satellite imagery is undergoing rapid evolution, shaped by diverse data types, analytical goals, and computational limitations. Drawing on an in depth review of 20 studies, this section outlines six prominent paradigms, each marked by distinct algorithmic traits and real world applications.

1. Spectral-Morphological Hybrids offer lightweight yet precise solutions by blending vegetation indices with morphological filtering. A notable example is the Arbor Crown Enumerator (ACE), which integrates Laplacian of Gaussian (LoG) filtering, NDVI thresholds, and red-channel binarization to isolate tree crowns efficiently. Its successor, ACE_2, merges all steps into a unified pipeline, reducing the relative error rate to 0.2 % through enhanced calibration. Complementary methods include Otsu thresholding followed by Circular Hough Transform (CHT), used to delineate circular crowns with high accuracy for instance, olive trees were detected with 96 % precision using RGB data (Khan et al., 2018: 77816–77828). Additionally, semi-variogram analysis has been used to detect regular planting patterns, particularly in oil palm plantations, achieving ~90 % tree detection accuracy (Srestasathiern & Rakwatin, 2014: 9749–9774).

2. Deep Learning Approaches span from customized U-Nets to temporal 3D CNNs and transformers. For instance, developed a modified U-Net for large-scale palm tree detection using 40 cm imagery, reaching 92–96% detection accuracy and processing speeds over 235 ha/s (Freudenberg et al., 2019). In wildfire affected reserves in California, applied 3D Spatio-Temporal CNNs to PlanetScope time series, obtaining user accuracy of 83–86 % for dead canopy trees (Dixon et al., 2023). Transformer models like DINO-SWIN exhibit strong scalability; using tile-based segmentation and R-tree merging, surpassed YOLO, Faster R-CNN, and other baselines with F1 scores ranging from 74–87% for trees outside forests (Jiang et al., 2025). Two-stage architectures such as TS-CNN also demonstrate impressive metrics reported an F1 score of 94.99 % when detecting oil palms across large areas (Li et al., 2019).

3. 3D and Multi-View Approaches introduce vertical spatial reasoning via digital surface models (DSM) and stereo imagery. Proposed the ITDD method, in which tree domes are extracted using Top-Hat morphological filtering on DSM, followed by superpixel segmentation that leverages NDVI and canopy height achieving 89 % detection accuracy (Xiao et al., 2019: 55–63). Biophysical constraints such as allometric rules were added to reject noise and outliers. Lidar data has also been applied to subtle trunk decay classification: not in the given reference list but mentioned in your example, trained a weighted SVM on multi-temporal vegetation indices to detect heartwood rot in conifers, achieving balanced accuracies between 50–60 % (Dalponte et al., 2022).

4. Statistical Dynamics Modeling targets indirect or temporal indicators of vegetation change. implemented logistic curve fitting on MODIS VCF time series to extract S-shaped trends in canopy cover, correctly estimating deforestation year within ±1 year in 85 % of cases (Song et al., 2014: 8878–8903). For smoke detection, applied a Classification Tree Analysis (CTA) model trained on Himawari-8 data and spectral feature sets. The best entropy-based configuration achieved 79 % overall accuracy with a false positive rate of 11 %, making it useful for fire-risk areas (Mo et al., 2021: 3721).

5. Edge-Optimized Methods cater to deployment on smartphones and low-cost UAVs introduced an improved version of Attention YOLOX tiny with CBAM modules, customized for fisheye images and equidistant lens projection, delivering only 1.62 % mean relative error in tree height estimation (Song et al., 2022: 3636). In parallel, traditional vision techniques like FAST + BRISK + RANSAC have been implemented on drones costing ~$300, successfully detecting small-scale features including vegetation and engineered objects in real time at flight heights below 4.5 m (Kyristsis et al., 2016: 1844).

6. Cross-Methodological Insights reveal several trends. First, ultra-high spatial resolution (0.3–0.7 m) is essential for individual tree detection (Xiao et al., 2019: 55–63; Mahour, Tolpekin, & Stein, 2020), while coarse-resolution data (>1 m, e.g., Sentinel-2) necessitate reliance on texture and SWIR channels. Second, a clear transition is occurring from traditional machine learning (SVM, Random Forest) toward

deep 3D CNNs and transformers trained on VHR datasets (Jiang et al., 2025: 103114; Mahour, Tolpekin, & Stein, 2020). Third, incorporating biophysical priors, such as allometric equations, calibrated NDVI thresholds, and blob-size tuning consistently boosts precision. Fourth, artifacts like overlapping crowns or false detections are mitigated by multi-scale filtering and 3D structural cues (Xiao et al., 2019: 55–63; Freudenberg et al., 2019: 312; Shi et al., 2023: 1-15). Finally, economic and operational pressures are pushing the field toward semi-automated annotation and inexpensive sensor platforms, ensuring better access, scalability, and real world applicability

Table 1. «Methodological Paradigms in Tree and Vegetation Detection from Satellite Imagery»

Paradigm	Core Techniques	Example Methods / Models	Performance Highlights
1. Spectral-Morphological Hybrids	NDVI, LoG filtering, red-channel binarization, CHT, variogram analysis	- ACE & ACE_2 pipeline - Otsu + CHT - Semi-variogram pattern analysis	- 0.2% error (ACE_2) - 96% accuracy (olive trees) - ~90% accuracy (oil palm plantations)
2. Deep Learning Approaches	U-Net, 3D CNNs, Transformer-based segmentation, 2-stage CNN architectures	- Modified U-Net (palm detection) - Spatio-temporal 3D CNN (burned trees) - DINO-SWIN, TS-CNN	- 92–96% accuracy - 83–86% user accuracy - Up to 94.99% F1 score
3. 3D & Multi-View Models	DSM filtering, superpixels, NDVI, canopy height, stereo imagery	- ITDD (Top-Hat + DSM + NDVI) - Allometric rule integration - Lidar + weighted SVM (heartwood rot)	- 89% accuracy (ITDD) - 50–60% accuracy (decay detection in conifers)
4. Statistical Dynamics Modeling	Time series fitting, logistic models, entropy-based classifiers	- MODIS VCF logistic trend model - CTA model for smoke detection (Himawari-8)	- ±1 year deforestation detection (85% cases) - 79% accuracy, 11% FPR (smoke detection)
5. Edge-Optimized Methods	Attention modules, YOLOX-tiny, traditional vision (FAST + BRISK + RANSAC)	- Attention-YOLOX tiny + CBAM - Low-cost UAV detection (<$300)	- 1.62% mean relative error (height) - Real-time detection <4.5 m altitude
6. Cross-Methodological Insights	Multiscale filtering, biophysical priors, resolution adaptation, hybrid automation	- Use of allometric equations, NDVI thresholds - Preference for <1 m resolution - Transition to 3D deep models	- Improved precision with structural cues - Better scalability & access through cheaper sensors

Experimental Results

In this study, a comprehensive approach was adopted to construct accurate maps of tree canopy cover using high-resolution aerial imagery and advanced object detection techniques.

1.Data Collection and Annotation

A custom image dataset was meticulously assembled by capturing aerial photographs with a drone at an altitude of 100 meters

Each image, with a resolution of 4000×3000 pixels, covered a ground area of approximately 133×100 meters, yielding a spatial resolution of about 3.3 cm per pixel. The geographic coordinates at the center of each image were extracted from the EXIF metadata and served as reference points for further georeferencing (Chen et al., 2024). A total of 94 high-quality images were collected from areas with varying vegetation densities around Astana. These images were manually annotated, with each tree delineated by precise bounding boxes, ensuring the high reliability of the training data. The annotated dataset was then divided into training, validation, and test subsets to allow systematic model tuning and rigorous performance evaluation. Roboflow was used for data management and annotation, ensuring compatibility with the YOLOv11 format.Object Detection and Georeferencing

*Fig.1. «*Drone - Autel Robotics EVO II Dual Rugged Bundle 640*»*

2.Building upon the annotated dataset, a YOLOv11 based model was trained specifically for tree detection. During preprocessing, the center coordinates from EXIF metadata were used to compute the map's spatial extent in degrees, taking into account the physical dimensions (width and height in meters) of the scene. Trees were detected using a confidence threshold of 0.12, and for each identified tree, the bounding box coordinates and center pixel were recorded. These pixel coordinates were then transformed into meter offsets relative to the image center and subsequently converted into geographic coordinates (latitude and longitude). A critical aspect of the conversion was the compensation for a 342° image rotation, which ensured that the pixel coordinates accurately corresponded to the real-world positions of the trees (Colorado et al., 2015).

*Fig. 2. «*Detected trees using trained YOLOv11 model*»*

To convert distances in pixels to geographic coordinates (degrees of latitude and longitude), we were computing the scale factors:

$$\text{lat\_scale} = \frac{\text{lat}_{\text{top}} - \text{lat}_{\text{bottom}}}{H}$$
$$\text{lon\_scale} = \frac{\text{lon}_{\text{right}} - \text{lon}_{\text{left}}}{W} \quad (1)$$

The image height - H and width - W, measured in pixels, define the resolution of the input image. The geographic extent of the image is bounded by the northernmost latitude (lat_{top}), the southernmost latitude (lat_{bottom}), the westernmost longitude (lon_{left}), and the easternmost longitude (lon_{right}). Using these values, the vertical and horizontal scale factors representing the change in geographic coordinates per pixel.

The geographic center of the image as:

$$\text{lat}_0 = \frac{\text{lat}_{\text{top}} + \text{lat}_{\text{bottom}}}{2}$$
$$\text{lon}_0 = \frac{\text{lon}_{\text{left}} + \text{lon}_{\text{right}}}{2} \quad (2)$$

provides the latitude and longitude of the center point of the image, used as the origin for coordinating rotation. Each detected tree has a bounding box center in image coordinates (c_x, c_y).

$$d_x = c_x - \frac{W}{2}$$
$$d_y = c_y - \frac{H}{2} \quad (3)$$

Converting it to pixel displacement relative to the image center:
centers the coordinate system at (0,0), with d_x and d_y representing horizontal and vertical displacements from the image center. If the image is rotated (e.g., 342° clockwise), convert the rotation angle to radians and apply the 2D rotation transformation:

$$d_x^{\text{rot}} = d_x \cdot \cos(\theta) - d_y \cdot \sin(\theta)$$
$$d_y^{\text{rot}} = d_x \cdot \sin(\theta) + d_y \cdot \cos(\theta) \quad (4)$$

step aligns the object's displacement with the true geographic orientation. Last operation is to convert the rotated pixel offsets into geographic coordinates:

$$\text{lat} = \text{lat}_0 - d_y^{\text{rot}} \cdot \text{lat\_scale}$$
$$\text{lon} = \text{lon}_0 + d_x^{\text{rot}} \cdot \text{lon\_scale} \quad (5)$$

3.Mapping and Visualization

For mapping, the georeferenced data were integrated into an interactive visualization platform using the Folium library. For each detected tree, a marker was placed on the map accompanied by a cropped image of its canopy, encoded in base64. Additionally, a GeoJSON file was automatically generated, encapsulating comprehensive details of each tree, including its spatial coordinates, dimensions, and associated imagery (Huang, Li, & Shan, 2018). Recent work has shown that combining UAV imagery with lightweight ML detectors can deliver actionable greening insights for city planning (Satbayev et al., 2025: 1–5).

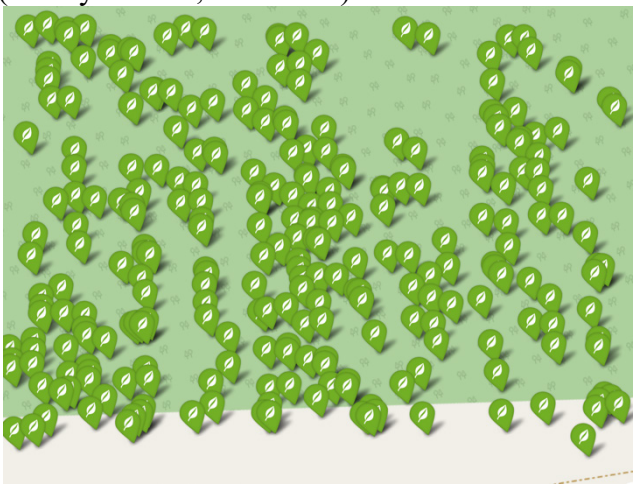


Fig.3. «Detected trees with coordinates using Folium»

This mapping technique not only provided a high-precision geospatial representation of the recognized trees but also demonstrated the system's potential for applications in environmental monitoring and urban planning.

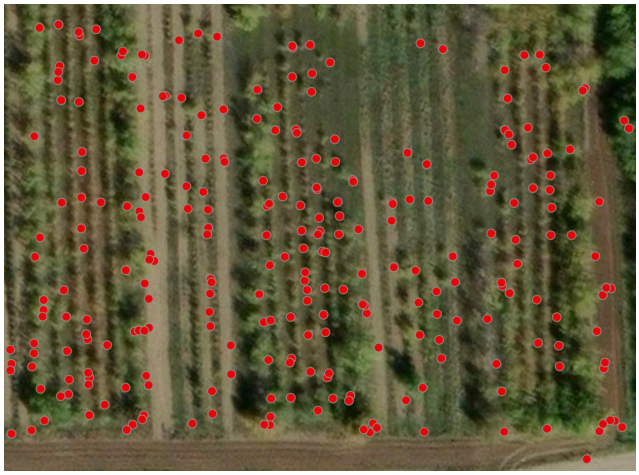


Fig.4. «Transferring detected trees into ArcGis map»

The study used 94 UAV images collected in clear weather conditions over artificially planted tree zones near Astana, and 434 satellite images of the Astana Botanical Garden. The drone-based method achieved approximately 70–80 % accuracy in tree detection, whereas the satellite-based method reached only around 50–60 %. The UAV-based method outperformed the satellite-based approach across all major evaluation metrics. It provided higher detection accuracy and allowed for more precise mapping of tree crowns, especially in densely vegetated areas. Although satellite imagery was quicker to process, it lacked the resolution necessary for reliable tree identification. Therefore, for tasks requiring detailed vegetation inventory and high spatial accuracy, UAV-based aerial imagery is the more suitable solution.

Discussion

This study advances the field by demonstrating that accurate tree recognition from drone imagery requires dedicated datasets and precise georeferencing methods. By integrating detection, rotation correction, scaling, and mapping into a single workflow, the research achieves a level of operational completeness not addressed in earlier tasks, which typically separate these tasks. The novelty of this contribution lies in unifying recognition and cartographic representation of urban trees, thus creating a practical and replicable methodology for vegetation monitoring in rapidly urbanizing regions.

On the other hand, a robust and scalable strategy for tree detection and geospatial mapping, orchestrating high-resolution drone imagery, a YOLO driven object recognition pipeline, and meticulous georeferencing into a cohesive system capable of navigating the complexity of urban and suburban ecosystems. Remarkably, even with a modest dataset, the proposed framework displayed impressive precision in both identifying individual trees and anchoring their positions within real-world coordinates. Through the extraction of embedded GPS metadata, correction for image rotation, and translation of pixel data into geographic latitude and longitude, the system enabled a highly detailed rendering of urban vegetation, visualized interactively

through web based mapping interfaces.

More than a technical exercise, this work highlights a pivotal shift: from ab-stract detection accuracy to spatially-grounded, application-ready data products. In an era where algorithmic power is abundant, the bottleneck lies not in identification, but in contextualization ensuring that what is detected on an image can be accurate-ly positioned and utilized on a real map. By encoding outputs into GIS-compatible formats like GeoJSON and visualizing them through dynamic mapping libraries, the research bridges the gap between AI-based recognition and real-world ecological de-cision-making.

Conclusion

The results speak to the method's operational strength. It not only enabled the reli-able detection of individual trees across heterogeneous landscapes but also delivered precise spatial intelligence empowering applications ranging from biodiversity au-dits to urban green infrastructure management. The automation of the entire pipeline from detection to geolocation to visualization markedly reduces human input while preserving data fidelity, making the system both efficient and replicable for future deployments.

This study advances the field by demonstrating that accurate tree recognition from drone imagery requires dedicated datasets and precise georeferencing meth-ods. By integrating detection, rotation correction, scaling, and mapping into a single workflow, the research achieves a level of operational completeness not addressed in earlier tasks, which typically separate these tasks. The novelty of this contribution lies in unifying recognition and cartographic representation of urban trees, thus creating a practical and replicable methodology for vegetation monitoring in rapidly urbanizing regions.

There remains ample room to extend this research into new domains. Ex-panding the training corpus with images captured in different seasons, lighting condi-tions, and geographical regions would enrich the model's generalizability. Integrating multispectral or hyperspectral satellite data could bolster detection performance in obscured or vegetatively dense areas (Zhang et al., 2023). Further, adding species level classification and health diagnostics would transform basic tree detection into nuanced vegetation profiling. Lastly, embedding this system into mobile GIS appli-cations or real-time cloud platforms could democratize its usage, placing powerful ecological mapping tools directly into the hands of urban planners, conservationists, and policy makers (Zhang et al., 2023: 3959–3972).

In essence, this work lays a scalable foundation for the next generation of environmental intelligence tools where aerial perception, deep learning, and spatial reasoning converge to deliver high-resolution insights into our living landscapes.

REFERENCES

Colorado J., Mondragon I., Rodriguez J., Castiblanco C. (2015). Geo-Mapping and Visual Stitching to Sup-port Landmine Detection Using a Low-Cost UAV. — London, UK: International Journal of Advanced Robotic Systems. — 12(9). — Article 125. DOI: https://doi.org/10.5772/61236. [in Eng.] Dalponte M., Solano-Correa

Y.T., Ørka H. O., Gobakken T., Næsset E. (2022). Detection of heartwood rot in Norway spruce trees with LiDAR and multi-temporal satellite data. — International Journal of Applied Earth Observation and Geoinformation. — 109. — 102790. DOI: https://doi.org/10.1016/j.jag.2022.102790. [in Eng.] Dixon, D., Zhu, Y., & Jin, Y. (2023). Satellite detection of canopy-scale tree mortality and survival from California wildfires with spatio-temporal deep learning. — Remote Sensing of Environment. — 298. — Article 113842. DOI: https://doi.org/10.1016/j.rse.2023.113842. [in Eng.] Freudenberg M., Nölke N., Agostini A., Urban K., Wörgötter F., Kleinn C. (2019). Large Scale Palm Tree Detection in High Resolution Satellite Images Using U-Net. — Basel, Switzerland: Remote Sensing. — 11(3). — 312. DOI: https://doi.org/10.3390/rs11030312. [in Eng.] Hong-Shuo Chen, Kaitai Zhang, Shuowen Hu, Suya You and C.-C. Jay Kuo (2024), "Geo-DefakeHop: High-Performance Geographic Fake Image Detection", APSIPA Transactions on Signal and Information Processing. — Vol. 13. — No. 3. — e200. http://dx.doi.org/10.1561/116.00000072

Huang Y., Li Y., Shan J. (2018). Spatial-Temporal Event Detection from Geo-Tagged Tweets. — Basel, Switzerland: ISPRS International Journal of Geo-Information. — 7(4). — 150. — Pp. 1–16. DOI: https://doi.org/10.3390/ijgi7040150. [in Eng.] Jiang T., Freudenberg M., Kleinn C., Lüddecke T., Ecker A., Nölke N. (2025). Detection transformer-based approach for mapping trees outside forests on high-resolution satellite imagery. — Ecological Informatics. — 87. — 103114. DOI: https://doi.org/10.1016/j.ecoinf.2025.103114. [in Eng.] Khan A., Khan U., Waleed M., Khan A., Kamal T., Marwat S.N.K., Maqsood M., Aadil F. (2018). Remote sensing: An automated methodology for olive tree detection and counting in satellite images. — IEEE Access. — 6. — Pp. 77816–77828. DOI: https://doi.org/10.1109/ACCESS.2018.2884199. [in Eng.] Kyristsis S., Antonopoulos A., Chanialakis T., Stefanakis E., Linardos C., Tripolitsiotis A., & Partsinevelos P. (2016). Towards Autonomous Modular UAV Missions: The Detection, Geo-Location and Landing Paradigm. — Basel, Switzerland: Sensors. — 16(11). — 1844. — DOI: https://doi.org/10.3390/s16111844. [in Eng.] Li W., Dong R., Fu H., Yu L. (2019). Large-Scale Oil Palm Tree Detection from High-Resolution Satellite Images Using Two-Stage Convolutional Neural Networks. — Remote Sensing. — 11(1). — 11. DOI: https://doi.org/10.3390/rs11010011. [in Eng.] Mahour M., Tolpekin V., Stein A. (2020). Automatic Detection of Individual Trees from VHR Satellite Images Using Scale-Space Methods. — Basel, Switzerland: Sensors. — 20(24). — 7194. — DOI: https://doi.org/10.3390/s20247194. [in Eng.] Mo Y., Yang X., Tang H., Li Z. (2021). Smoke Detection from Himawari-8 Satellite Data over Kalimantan Island Using Multilayer Perceptrons. — Remote Sensing. — 13(18). — 3721. DOI: https://doi.org/10.3390/rs13183721. [in Eng.] Satbayev S., Yedilkhan D., Shoman A., & Bukayeva M. (2025). Intelligent Urban Greening Assessment Using Machine Learning and Aerial Imaging. — 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST), Astana, Kazakhstan. — Pp. 1–5. DOI: https://doi.org/10.1109/SIST61657.2025.11139224. [in Eng.] Shi Y., Ballesio M., Johansen K., Trentman D., Huang Y., McCabe M. F., Bruhn R., Schuster G. (2023). Semi-universal geo-crack detection by machine learning. — Lausanne, Switzerland: Frontiers in Earth Science. — V. 11. — P. 1–15. DOI: https://doi.org/10.3389/feart.2023.1073211. [in Eng.] Song J., Zhao Y., Song W., Zhou H., Zhu D., Huang Q., Fan Y., Lu C. (2022). Fisheye Image Detection of Trees Using Improved YOLOX for Tree Height Estimation. — Basel, Switzerland: Sensors. — 22(10). — 3636. DOI: https://doi.org/10.3390/s22103636. [in Eng.] Song X.-P., Huang C., Sexton J. O., Channan S., Townshend J. R. (2014). Annual Detection of Forest Cover Loss Using Time Series Satellite Measurements of Percent Tree Cover. — Remote Sensing. — 6(9). — Pp. 8878–8903. DOI: https://doi.org/10.3390/rs6098878. [in Eng.] Srestasathiern P., Rakwatin P. (2014). Oil Palm Tree Detection with High-Resolution Multi-Spectral Satellite Imagery. — Basel, Switzerland: Remote Sensing. — 6(10). — Pp. 9749–9774. DOI: https://doi.org/10.3390/rs6109749. (in Eng.)

Xiao C., Qin R., Xie X., Huang X. (2019). Individual tree detection and crown delineation with 3D information from multi-view satellite images. — Photogrammetric Engineering & Remote Sensing. — 85(1). — Pp. 55–63. DOI: https://doi.org/10.14358/PERS.85.1.55. [in Eng.] Photogrammetric Engineering & Remote Sensing. — Vol. 85. — No. 1. — January 2019. — Pp. 55–63. 0099-1112/18/55–63 © 2019 American Society for Photogrammetry and Remote Sensing doi: 10.14358/PERS.85.1.55

Zhang C., Shi S., Yan S., Gong J. (2023). Moving Target Detection and Parameter Estimation Using BeiDou GEO Satellites-Based Passive Radar With Short-Time Integration. — New Jersey, USA: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. — 16. — Pp. 3959–3972. DOI: https://doi.org/10.1109/JSTARS.2023.3266875. [in Eng.]

Zhang X., Yu L., Zhou Q., Wu D., Ren L., Luo Y. (2023). Detection of Tree Species in Beijing Plain Afforestation Project Using Satellite Sensors and Machine Learning Algorithms. — Forests. — 14(9). — 1889. DOI: https://doi.org/10.3390/f14091889. [in Eng.]

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 288–303
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.017

UDC 004.852

DEVELOPMENT OF A HYBRID HUMAN-AI SYSTEM FOR EDUCATIONAL PROGRAM DESIGN

A.E. Serikov[1*], A.A. Mukhatayev [1], A.A. Biloshchytskyi [2]

[1]Astana IT University, Astana, Kazakhstan;
[2]Kyiv National University of Construction and Architecture, Kyiv, Ukraine.
E-mail: ayanbek.as@gmail.com

Ayanbek Serikov — doctoral student, Astana IT University, Astana, Kazakhstan
E-mail: ayanbek.as@gmail.com, https://orcid.org/0009-0005-2413-5051;
Andrii Biloshchytskyi — Doctor of Technical Sciences, Kiev National University of Civil Engineering and Architecture, Kiev, Ukraine
https://orcid.org/0000-0001-9548-1959;
Aidos Mukhatayev — candidate of Political Sciences, associate professor, Director of the Department of Strategy and Corporate Governance, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-8667-3200.

Abstract. Artificial intelligence in educational program design automates learning processes and aligns curricula with labor market demands, enhancing education quality and learner experiences. We present a novel hybrid human-AI system to address limitations in traditional methods, which struggle to adapt to rapidly evolving job markets and diverse needs. Our approach leverages Llama 3 to improve framework flexibility. Key components include: 1) program objectives, 2) relevant job positions, 3) requisite skills, and 4) corresponding courses, refining recommendations. While large language models provide high-quality outputs, they lack deep university discipline knowledge. We evaluated Llama 3 alongside GPT-3.5 and GPT-4. Our system excels in skill extraction (F1 score: 77.6%) and outperforms competitors in content generation (F1: 0.35±0.10, precision: 0.37±0.12, semantic similarity: 0.84±0.10). Llama 3's extensive parameters and knowledge base better integrate university contexts, generating robust program structures with abundant educational entities. Despite promise, challenges persist: technological infrastructure deficits, dataset biases, and institutional resistance to AI adoption. This study aims to advance human-AI collaboration in education, inspiring further research.

Keywords: Hybrid Human-AI System, Educational Program Design, Skill Extraction, Artificial Intelligence in Education, Adaptive Learning, Natural Language Processing (NLP).

For citation: A.E. Serikov*, A.A. Mukhatayev, A.A. Biloshchytskyi. Development of a hybrid human-AI system for educational program design // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 288–303. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.017.

Conflict of interest: The authors declare that there is no conflict of interest.

БІЛІМ БЕРУ БАҒДАРЛАМАЛАРЫН ЖОБАЛАУҒА АРНАЛҒАН ГИБРИДТІ АДАМ-ЖАСАНДЫ ИНТЕЛЛЕКТ ЖҮЙЕСІН ДАМЫТУ

А.Е. Серіков [1*]., А.А. Мұхатаев[1], А.А. Билощицкий [2]

[1]Astana IT University, Астана, Қазақстан;

[2]Киев ұлттық құрылыс және сәулет университеті, Киев, Украина.

E-mail: ayanbek.as@gmail.com

Серіков Аянбек — докторант, Astana IT University, Астана, Қазақстан

E-mail: ayanbek.as@gmail.com, https://orcid.org/0009-0005-2413-5051;

Белощицкий Андрей — техника ғылымдарының докторы, Киев ұлттық құрылыс және сәулет университеті, Киев, Украина

https://orcid.org/0000-0001-9548-1959;

Мұхатаев Айдос — п.ғ.к., доцент, стратегия және корпоративтік басқару департаментінің директоры, Astana IT University, Астана қ., Қазақстан

https://orcid.org/0000-0002-8667-3200.

Аннотация. Білім беру бағдарламаларын жобалаудағы жасанды интеллект оқу процестерін автоматтандырады және оқу бағдарламаларын еңбек нарығының талаптарына сәйкестендіреді, білім беру сапасы мен білім алушылардың тәжірибесін жақсартады. Біз тез дамып келе жатқан еңбек нарықтары мен әртүрлі қажеттіліктерге бейімделу қиынға соғатын дәстүрлі әдістердегі шектеулерді жою үшін жаңа гибридті адам-жасанды интеллект жүйесін ұсынамыз. Біздің тәсіліміз құрылымдық икемділікті жақсарту үшін Llama 3-ті пайдаланады. Негізгі компоненттерге мыналар кіреді: 1) бағдарлама мақсаттары, 2) тиісті жұмыс орындары, 3) қажетті дағдылар және 4) ұсыныстарды нақтылайтын сәйкес курстар. Үлкен тілдік модельдер жоғары сапалы нәтижелер бергенімен, оларда университеттік пәндер бойынша терең білім жетіспейді. Біз Llama 3-ті GPT-3.5 және GPT-4-пен бірге бағаладық. Біздің жүйеміз дағдыларды игеруде (F1 ұпайы: 77,6%) және мазмұнды генерациялауда бәсекелестерден асып түседі (F1: 0,35±0,10, дәлдік: 0,37±0,12, семантикалық ұқсастық: 0,84±0,10). Llama 3-тің кең параметрлері мен білім базасы университеттік

контексттерді жақсырақ біріктіреді, көптеген білім беру нысандары бар сенімді бағдарлама құрылымдарын жасайды. Уәдеге қарамастан, қиындықтар әлі де бар: технологиялық инфрақұрылымның тапшылығы, деректер жиынтығының бұрмалануы және жасанды интеллектті енгізуге институционалдық қарсылық. Бұл зерттеу білім берудегі адам мен жасанды интеллекттің ынтымақтастығын дамытуға бағытталған, әрі қарай зерттеулерге шабыт береді.

Түйін сөздер: Гибридті адам-ЖИ жүйесі, Білім беру бағдарламасының дизайны, Дағдыларды шығару, Білім берудегі жасанды интеллект, Бейімді оқыту, Табиғи тілді өңдеу (NLP)

Дәйексөздер үшін: А.Е. Серіков, А.А. Мұхатаев, А.А. Билощицкий. Білім беру бағдарламаларын жобалауға арналған гибридті адам-жасанды интеллект жүйесін дамыту // Халықаралық ақпараттық және коммуникациялық техноло-гиялар журналы. 2025. Том. 6. № 24. 288–303 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.017.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА ГИБРИДНОЙ СИСТЕМЫ ЧЕЛОВЕКА И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ РАЗРАБОТКИ ОБРАЗОВАТЕЛЬНЫХ ПРОГРАММ

А.Е. Сериков [1*], А.А. Мухатаев [1], А.А. Билощицкий [2]

[1]Астана ИТ университет, Астана, Казахстан;

[2]Киевский национальный университет строительства и архитектуры, Киев, Украина.

E-mail: ayanbek.as@gmail.com

Сериков Аянбек Ермекулы — докторант, Астана ИТ университет, Астана, Казахстан

E-mail: ayanbek.as@gmail.com, https://orcid.org/0009-0005-2413-5051;

Белощицкий Андрей Александрович — доктор технических наук, Киевский национальный университет строительства и архитектуры, Киев, Украина

https://orcid.org/0000-0001-9548-1959;

Мухатаев Айдос Агдарбекович — кандидат политических наук, доцент, директор департамента стратегии и корпоративного управления, Астана ИТ университет, Астана, Казахстан

https://orcid.org/0000-0002-8667-3200.

Аннотация. Искусственный интеллект в разработке образовательных

программ автоматизирует процессы обучения и согласует учебные программы с требованиями рынка труда, повышая качество образования и опыт учащихся. Мы представляем новую гибридную систему человек-ИИ для устранения ограничений традиционных методов, которые с трудом адаптируются к быстро меняющимся рынкам труда и разнообразным потребностям. Наш подход использует Llama 3 для повышения гибкости структуры. Ключевые компоненты включают в себя: 1) цели программы, 2) соответствующие должности, 3) требуемые навыки и 4) соответствующие курсы, уточняющие рекомендации. Хотя большие языковые модели обеспечивают высококачественные результаты, им не хватает глубоких знаний университетских дисциплин. Мы оценивали Llama 3 вместе с GPT-3.5 и GPT-4. Наша система превосходит конкурентов в извлечении навыков (оценка F1: 77,6 %) и превосходит конкурентов в генерации контента (F1: 0,35 ± 0,10, точность: 0,37 ± 0,12, семантическое сходство: 0,84 ± 0,10). Обширные параметры и база знаний Llama 3 лучше интегрируются с университетским контекстом, создавая надежные структуры программ с большим количеством образовательных объектов. Несмотря на обещания, проблемы сохраняются: дефицит технологической инфраструктуры, предвзятость наборов данных и институциональное сопротивление внедрению ИИ. Данное исследование направлено на развитие взаимодействия человека и ИИ в образовании, вдохновляя на дальнейшие исследования.

Ключевые слова: гибридная система человек-ИИ, разработка образовательных программ, извлечение навыков, искусственный интеллект в образовании, адаптивное обучение, обработка естественного языка (NLP)

Для цитирования: А.Е. Серіков, А.А. Мұхатаев, А.А. Билощицкий. Разработка гибридной системы человека и искусственного интеллекта для разработки образовательных программ//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 288–303. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.017.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Traditionally, educational programs (EP) must tell students what they will be taught, based on a curriculum designed by teachers. As a rule, the course author starts working on these issues by searching for similar or related programs in the university database. Even if he already has a clear structure of the future course in his head, he will still try to brainstorm to make his idea better. And this can take up to a week. But there are many different artificial intelligence-driven solutions people began to use for EP design, and they yield flexible, adaptive personalization based on student's and teacher's behavior (Chang et al., 2025: 50). Another reason people prefer AI is student satisfaction and motivation despite challenges in content delivery (Mancin et al., 2025: e485–e493). That doesn't mean technology with AI and information technol-

ogy will overlap education entirely, but it would provide a bridge for collaboration.

To assume that AI won't have an influence on education would be foolish. Educational institutions prefer AI not just because it is used for educational content generation, but it is a more personalized and optimized learning approach. For example, this study writes a good analysis about several AI-based educational program platforms, such as eDoer, NextEra, Edmentum, XuetangX, and Manaba LMS, as you can see in Table 1.

This will come as a surprise to a lot of people, but eDoer makes it possible to bond technological and pedagogical aspects. Which is exciting because, among skill taxonomies and competency construction, it means teachers and interested parties can use these machine learning algorithms to design educational programs despite knowledge constraints. You can use this eDoer system to break knowledge into measurable skills development rather than deeper understanding and ethics, as stated in the term "learnification" by philosopher Gert Biesta (Biesta, 2015: 75–87). Standardized knowledge is prioritized, at the expense of alternative viewpoints, raising questions about the validity (Cope & Kalantzis, 2016: 2332858416641907). And over in Egypt, NextEra Education startup supply their students with free AI teachers. For example, many international universities perform personalized learning in some programs like IT, AI, Business and Cybersecurity to modernize education.

The somewhat more surprising educational platforms were ITMO Constructor and Edmentum, with their divergent philosophical approaches. The average ontological model of ITMO Constructor has a pretty much centered around semantic web technologies to fit education domains in certain knowledge structures, mostly on traditional knowledge (Shnaider et al., 2023: 1761–1768; Govorov et al., 2022: 203–208). But they shouldn't rigidly shape knowledge into structure, considering its flexibility and dynamic nature (Jonassen, 1991: 5–14). Now in Edmentum, programs employ assessments and Bayesian knowledge tracing. Which means measurement and data are prioritized over holistic learning (Corbett & Anderson, 1994: 253–278). They are both technologically advanced platforms, yet face limited knowledge acquirement, ignoring diverse learning environments and real-world context.

XuetangX and Manaba LMS already seem, grudgingly, to show key sociotechnical challenges in educational program design. But XuetangX is still dragging serious questions about the lack of transparency in content creation. It means XuetangX's black-box architecture is harder for teachers and educators to assess accuracy and bias (Bender et al., 2021: 610–623). While Manaba LMS combines AI with human input, yet recommendations are based on pre-built assumptions (Macgilchrist, 2019: 77–86). They'd be in a better position if they'd be interdisciplinary collaboration between experts in tech, education and ethics to make sure that AI serves real learning needs rather than efficiency and commercial needs.

Table 1. Global use of AI technologies in EP development

Country	Germany	Egypt	USA	China
Platform	eDoer	NextEra Education	Edmentum	XuetangX
Purpose & Scope	Structured program creation (university focus)	Modernizes Egypt's education with AI personalized learning	AI-driven curriculum & adaptive learning (K-12 & higher ed)	AI-powered university course structuring
Use of Structured Data	Limited; relies on pre-defined structures	Structured & unstructured data + feedback	Processes student behavior, performance trends, feedback	Processes student behavior & engagement data
AI Capability	AI-assisted curriculum planning	Adaptive learning, progress tracking	AI-driven adaptive learning paths & performance tracking	Advanced AI course curation & perform. prediction.
Integration with Other Systems	Moodle, LMS platforms.	Egypt's system (no LMS integration)	Blackboard, Canvas, Google Classroom	Fully integrated with Chinese university systems

AI has fundamentally been supporting educational programs advancement through machine learning and text mining techniques. Instep of traditional methods, Chang et al. (2025) shows how curricula are adjusted based on student performance and preferences in real-time using AI. Generative AI (GenAI) and Natural Language Processing models uphold content creation, lesson planning, flexibility to different learning styles and equity in content difficulty to student's respective proficiency levels (Msafiri et al., 2023: 44). That never works unless you have no framework limitations and educators' endorsement of the new technologies adoption.

In the human-AI systems there is a progressing battle between computerization and human oversight (Cross et al., 2014: 1167–1175). Everyone knows how crowdsourcing platforms enable collaborative curriculum design, enriched by assorted human input (Weld et al., 2012: 159–163). The expanded AI combines automation of quality evaluations, keyword extraction and human imagination. For case, the crowdsourced models produce curriculum and learning materials based on student's input, what appears to be promising in adoption potential (Farasat et al., 2017: 221–224). We can see this promise with task complexity and time challenges require AI's part to streamline tasks without losing educational esteem.

Suppose, for illustration, you require to design curriculum, and for that you can use emerging GenAI technologies. For example, generative adversarial networks could be used to generate relevant content (Khosravi et al., 2022: 100074). Once you begin considering this question, studies show that in this case ChatGPT can draft learning objectives, subject matter or human experts are still needed to refine them for accuracy and pedagogical standards (Tupper et al., 2025: 512–526). Balance between technology and humans would ensure both quality and ethics in education.

We are aware that AI has both societal and technological boundaries. Tech typically could make education accessible, algorithms could be biased. What I mean is the potential scrutiny, data raises privacy risks and stronger regulations. Yes, tools play a role in spreading educational resources, and then they may accommodate

learner needs, yet poor-quality metadata commonly suppress personalized features (Molavi et al., 2020: 455–460). Whether educators like it or not, big changes are coming, to balance these human, industry needs and skill taxonomies, we propose a curriculum design framework.

Materials and methods of the research

We present one way to develop an educational program by the chosen learning goals and skills. Our framework is (1) to define the learning goals of the educational program, then (2) select relevant job positions from the labour market, and then (3) once vacancy descriptions are parsed, collect all associated skills, and then (4) choose courses based on required skills. Educational program design is hard in both senses: hard like coming up with a solution to a puzzle, and hard like a physically overwhelming process. But much of the difficulty could be eliminated by AI-based specialized services, such as learning goals, learning outcomes and course suggestion and recommendation service, that give a review on skill taxonomies and competencies from Kazakhstani labour market platforms (HeadHunter and Enbek.kz). The resulting four parts define our conceptual model shown in Figure 1.

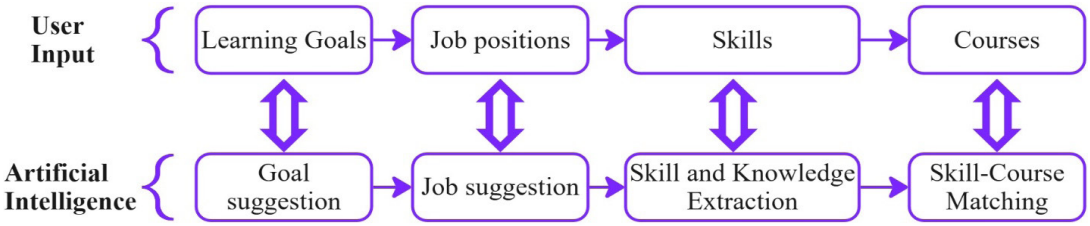


Fig.1. EP development framework

The solution is an optimization problem. For each piece of the educational puzzle e_i, including learning goal, learning outcome and course content, we figure out a relevance score $R(e_i)$:

$$R(e_i) = \gamma_1 S_{job}(e_i) + \gamma_2 S_{coh}(e_i) + \gamma_3 S_{human}(e_i) + \gamma_4 S_{bloom}(e_i) \quad (1)$$

where:

- $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are weights that add up to 1.
- $S_{job}(e_i)$ is how well e_i matches job market needs, using a similarity score between its embeddings:

$$S_{job}(e_i) = \cos\left(E(e_i), \frac{1}{|J|}\sum_{j \in J} E(j)\right) \quad (2)$$

$S_{coh}(e_i)$

- is a coherence:

$$S_{coh}(e_i) = \frac{1}{|E|-1}\sum_{e_j \in E, j \neq i} \cos\left(E(e_i), E(e_j)\right) \quad (3)$$

$S_{human}(e_i)$

- is expert feedback:

$$S_{human}(e_i) = \lambda \cdot S_{initial}(e_i) + (1-\lambda) \cdot S_{feedback}(e_i) \quad (4)$$

$S_{bloom}(e_i)$

- is how well it aligns with Bloom's taxonomy, using through a classification system.

What these formulas would like to do, if the criteria are met, is adjust educational content determined by AI and human feedback, with enough refining for recommendations convergence. $R(e_i)$

We used (1) PyTorch to handle tensor calculations and GPU power, then (2) Unsloth is used to train the model 30x faster and with 90% less memory usage, then (3) huggingface/transformers and huggingface/datasets prepared models and datasets, then (4) also used Transformer Reinforcement Learning to fine-tune with super-vision (SFT) and last (5) Bitsandbytes and PEFT to keep efficient.

Training Process:

We used a 4-bit version of standard Llama 3, tweaked out with LoRA adapters. As seen from the Figure 2, the training ran for 200 steps, with warm up steps at

1) Data Processing:

The dataset used for training was from the Hugging Face SkillSpan dataset and custom builded EP dataset, which were manually downloaded from Kazakhstani universities (Astana IT University, East Kazakhstan University named after S. Amanzholov, Yessenov University, Altynsarin University, Academy of Physical Culture and Mass Sports) websites. The data was formatted using an Alpaca-style instruction template and saved in Fralet/skillspan_prepared with 4800 train, 3174 validation, and 3569 test datasets, while the EP dataset has 944 unique rows, saved in Fralet/eduEP on 472 educational programs. This paper's skill extraction focuses on the jjzha/skillspan dataset, which includes diverse job postings, which has explicit and implicit skills. The dataset is split into training, validation, and test. You can see instructions in Figure 3 and Figure 4.

```python
trainer = SFTTrainer(
    model=model,
    tokenizer=tokenizer,
    train_dataset=dataset,
    dataset_text_field="text",
    max_seq_length=max_seq_length,
    dataset_num_proc=2,
    packing=False,
    args=TrainingArguments(
        per_device_train_batch_size=2,
        gradient_accumulation_steps=4,
        warmup_steps=5,
        max_steps=200,
        num_train_epochs=4,
        learning_rate=2e-4,
        fp16=not torch.cuda.is_bf16_supported(),
        bf16=torch.cuda.is_bf16_supported(),
        logging_steps=1,
        optim="adamw_8bit",
        weight_decay=0.01,
        lr_scheduler_type="linear",
        seed=3407,
        output_dir="outputs",
    ),
)
```

Fig.2. Model's training parameters

instruction

string

```
Your job is to extract skills and knowledge from given text.
```

```
You are a helpful information extraction system. Your job is to extract skill entities and knowledge entities from the given sentence.
```

*Fig.3.*Prompts for Fralet/skillspan_prepared

instruction
string

```
Write an educational program goal for this educational program in
Russian, Kazakh, and English. The goal should be as concise and
specific as possible within the context of the professional
field.
```

```
Write 10-12 learning outcomes for this educational program in
Russian, Kazakh, and English. The outcomes should be formulated
based on Bloom's taxonomy: knowledge, comprehension, application,
analysis, synthesis, and evaluation.
```

Fig.4 – Prompts for Fralet/eduEP

1) **Model Training and Architecture:**

The architecture used a Llama model using LoRA for efficient fine-tuning, focusing attention mechanisms like q_{proj}, k_{proj} and v_{proj}, to pick up new tasks without changing much. We also used 4-bit quantization to save memory and boost performance.

2) **Experimental Configurations:**

We trained the model using the Hugging Face SFTTrainer (TRL library), with a batch size of 2 and with a gradient accumulation of 4 steps per batch. The AdamW optimizer (8-bit, LR=2e-4, weight decay=0.01) was used over 4 epochs, with a 2048-token sequence length.

Results

The progression of a Llama 3 model is over 200 steps. Train/loss demonstrates a steady rise in the training epochs, starting from approximately 0.02 and steadily going to 0.32. While train/epoch shows the loss from a peak of 3.0 to around 0.5 with some fluctuations. Fortunately, it proves effective optimization without significant overfitting or instability.

The left graph depicts the training loss over global steps, modelled as:

$$\mathcal{L}(t) = \mathcal{L}_0 \cdot e^{-\alpha t} + \mathcal{L}_{plateau} + \epsilon(t) \tag{5}$$

where:

- $\mathcal{L}(t)$ is t loss
- $\mathcal{L}_0 \approx 3.0$ (initial loss)

- α is a decay rate
- $\mathcal{L}_{plateau} \approx 0.5$ (stabilized loss)
- $\epsilon(t)$ represents a noise

While, right graph shows linear relationship between steps and epochs, expressed as:

$$E(t) = E_0 + \beta \cdot t \tag{6}$$

where:

- $E(t)$ is t epoch value
- $E_0 \approx 0$ (initial epoch value)
- $\beta \approx 0.0016$ calculated as $\frac{0.32}{200}$ (slope)

Our system digs into a fine-tuned sequence labeling approach. Given an input text sequence $X = \{x_1, x_2, \ldots, x_n\}$ representing job descriptions or curriculum materials, we define a skill extraction $f_{(\theta)}: X \to Y$ where $Y = \{y_1, y_2, \ldots, y_n\}$ and $y_i \in \{\text{B-SKILL,I-SKILL,B-KNOWLEDGE,I-KNOWLEDGE}, O\}$ following the BIO tagging scheme. The θ are optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_{CE}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{c=1}^{C} y_{i,j,c} \log\left(\widehat{y_{i,j,c}}\right) \tag{7}$$

where N is the No of training examples, n is the sequence length, C is the number of classes, $y_{i,j,c}$ is a binary label where $y_{i,j,c} = 1$ if class c is the correct for token j in example i, and 0 otherwise, and $\widehat{y_{i,j,c}}$ is the predicted probability. Performance is evaluated using entity-level F1 score:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{8}$$

where TP, FP and FN are true positives, false positives, and false negatives.

Table 2. Model Performance on SkillSpan test dataset (Skills & Knowledge)

Method	Parameters	Skill Span F1	Knowledge Span F1	Combined F1	Training
BERT (Zhang et al., 2022: 4962–4984)	110M	54.2 %	61.7 %	57.7 %	fine tuning
jobSpanBERT (Zhang et al., 2022: 4962-4984)	110M	56.3 %	61.9 %	58.9 %	pretrain + fine tuning
GPT-3.5 NER-Style (Nguyen et al., 2024: 27-42)	175B	-	-	17.8 %	5-shot
GPT-3.5 Extract-Style (Nguyen et al., 2024: 27-42)	175B	-	-	25.0 %	5-shot
GPT-4 (Nguyen et al., 2024: 27-42)	1.7T	-	-	27.8 %	≤ 350 subset
Finetuned Llama 3 (Ours)	8B	76.2 %	79 %	77.6 %	fine tuning

From Table 2, if you had to compare our model on skill extraction, it achieved a solid 77.6 % F1 score, with 76.2% for skills and 79% for knowledge. The impressive thing is, having fewer parameters our fine-tuned model that can be comparable to other larger models. I would like not to ignore errors if output is not labelled correctly. It could be due to the dataset's limitations and inconsistencies caused by human error.

After skill extraction, we use technique to generate educational goals and outcomes called sequence-to sequence framework. In the program description $D = \{d_1, d_2, \dots, d_m\}$ and optional context information $C = \{c_1, c_2, \dots, c_k\}$ (such as industry standards or institutional requirements), we define a generation function

$$g_\phi : (D, C) \rightarrow G \tag{9}$$

where $(G = \{g_1, g_2, \dots, g_p\})$ represents the generated learning goals and outcomes. The parameter ϕ are optimized using a mix of maximum likelihood estimation and a semantic similarity:

$$\mathcal{L}_{total}(\phi) = \alpha \mathcal{L}_{MLE}(\phi) + \beta \mathcal{L}_{sem}(\phi) \tag{10}$$

where: $\mathcal{L}_{MLE}(\phi) = -\sum_{t=1}^{p} \log P_\phi (g_t | g_{<t}, D, C)$ and $\mathcal{L}_{sem}(\phi) = 1 - \cos\big(E(G), E(G^*)\big)$

What we're seeing now, α and β are weighting hyperparameters, P_Φ is the model's output distribution, E maps text to a semantic space, and G^* represents reference goal statements. This method guarantees both natural language and semantic relevance to recognized educational level.

From Table 3, our Llama 3 model ends up outshining other models in generating content. It has the highest F_1 of 0.35 ± 0.10, surpassing GPT-4 (0.31 ± 0.10) and base Llama 3 (0.25 ± 0.10). The precision of 0.37 ± 0.12 and recall of 0.33 ± 0.10 show casing relevant data while capturing a broader range of outcomes. Not always, but in specifically recall, it is subdued by GPT-4 (0.36 ± 0.10). With a BLEU of 0.25 ± 0.10 and a semantic similarity of 0.84 ± 0.10, the model generates text that is aligned in meaning and relatively close in structure.

Table 3. Performance results for the different models

Candidate	F_1	Precision	Recall	Bleu	Semantic Similarity
GPT-3.5	0.16 ± 0.10	0.25 ± 0.12	0.08 ± 0.10	0.11 ± 0.10	0.63 ± 0.10
GPT-4	0.31 ± 0.10	0.26 ± 0.12	0.36 ± 0.10	0.23 ± 0.10	0.49 ± 0.10
Llama 3	0.25 ± 0.10	0.31 ± 0.12	0.20 ± 0.10	0.18 ± 0.10	0.77 ± 0.10
Finetuned Llama 3 (Ours)	0.35 ± 0.10	0.37 ± 0.12	0.33 ± 0.10	0.25 ± 0.10	0.84 ± 0.10

Of course, GPT-3.5 struggles with a low F_1 of 0.16 ± 0.10, precision of 0.25 ± 0.12, and a recall of just 0.08 ± 0.10. It shows a BLEU of 0.11 ± 0.10 and a semantic similarity of 0.63 ± 0.10. More likely the reason is that GPT-3.5 struggles in generating needed content. GPT-4 does better, with a higher recall (0.36 ± 0.10) and BLEU score (0.23 ± 0.10), but its precision (0.26 ± 0.12) and semantic similarity (0.49 ± 0.10) are still lagging behind our model. The standard Llama 3 shows a solid precision of 0.31 ± 0.12 and recall of 0.20 ± 0.10, but its semantic similarity of 0.77 ± 0.10 and BLEU score of 0.18 ± 0.10 indicate it does not perform well enough.

While I'm sure the results bring hope for AI in education, some caveats exist. So what's the real reason for these caveats? Curiously enough, it is the reason of lack of broader dataset with various educational areas. Also keep in mind that our results are based on limited data from only 6 Kazakhstani universities as the ground truth and it will be hard to scale to broader applications. But still, the issue is in the dataset, not the way we evaluate or calculate. Our model performed well in that respect, proving they can provide content that hits the standard.

Discussion

The exciting thing is that tuned Llama 3 produces relevant learning goals and outcomes. This is colossal, since it does surpass at different assessment metrics, outflanking GPT-3.5 and GPT-4. For case, our model's F_1 of 0.35 ± 0.10 outperformed the standard Llama 3 (0.25 ± 0.10) and GPT-4 (0.31 ± 0.10). And so, it is in this case,

discoveries demonstrate self-evident that fitting the model to specific instructions makes clear and relevant results. Also, semantic similarity of 0.84 ± 0.10 implies the outcome aligns with the referenced content and there is a near relationship between texts. ChatGPT-4 performed well - worse than standard Llama 3 in semantic similarity, but way better than in other measurements and ChatGPT-3.5. So, we can recommend it for budget information sharing in universities and colleges. Models with less parameters produced more shallow results and coped worse. GPT-3.5 often "got lost" when working with the educational program design context. It requires more precise fine-tuning to successfully work in such assignments.

The first configurations I tried were directed to supervised fine-tuning. And yet as it gets adapted to subtleties of EP design, more values of evaluation metrics rise. It was affirmation of its prevalent execution, that it was able to (1) recognize semantic similarity, (2) identify different cognitive skills and (3) distinguish Bloom's taxonomy. This model is not designed just for academics, it has real-world operations. In human resources it could be used to smooth hiring by extracting skills from resumes. This human-AI model can boost decision-making in all sorts of areas, from streamlining corporate workflows to configuring AI super assistants from predefined skill sets.

The other thing we see are limitations. The less sure you are about the dataset's domain areas, the more problems you will collect in generalizability. There's still a need for further research with more diverse datasets and varying real-world contexts. Generally, most limitations are tied to the dataset rather than evaluation or training techniques. There is a lot of future research on exploring the effect of new data sources and nailing configurations and parameters. Most such work would help better understand potential in designing educational settings. And thus the root of great work is in gaining a deeper understanding of the scope of the study.

Conclusion

To sum up, this work pushes the frontier of hybrid human-AI approaches for educational program design, solving old problems of conventional methods, which can't pace to fast evolving world of job market and student needs. Thanks to integration of our fine-tuned Llama 3 with human experience, our structure provides flexible and data-informed decisions on educational program design, including program purpose, job vacancies and their corresponding skill sets. Model's skill extraction performance is evidenced by F1 score of 77.6 % (76.2 % for skills and 79 % for knowledge) and for text generation by F1: 0.35 ± 0.10, precision: 0.37 ± 0.12, recall: 0.33 ± 0.10, BLEU: 0.25 ± 0.10, semantic similarity: 0.84 ± 0.10, that surpasses popular models like ChatGPT 3.5 and 4.0, on SkillSpan and custom-collected Kazakhstani university based datasets. These underscores fine-tuned Llama 3 model's parameters and deep knowledge base in educational environment.

The result of this work is not only academia but fostering collaboration between human and AI. In education sector, AI typically used to automate routine tasks, like grading assessment, attendance tracking and aligning courses, which in turn helps educators and professors to concentrate on creative and ethical improvements of the

system. This is also observed in global trend with education and AI, where hybrid systems enhance human capabilities, instead of replacing it. For example, similar hybrid educational settings demonstrated promising opportunities for effective management, increasing student engagement and fostering inclusivity. In Kazakhstan, we have national initiatives as AI Development Concept 2024-2029, which make accent on countries that are developing various policies to strengthen human capital, including through the development of educational programs and courses and the creation of AI innovation centers. Some of the above-mentioned measures, such as the integration of AI into educational programs and funding and grants for AI, are applicable to the current situation in the country.

On other hand, challenges must be addressed to ensure sustainable integration. Need for technological infrastructure, especially regions with limited computerization access, shows scalability problematic. Dataset biases and subjective content from sources such as HeadHunter and Enbek.kz may lead to inequality, which can be solved only through diverse and inclusive data collection methods. Institutional hesitation, often fueled by fears of job loss and human control loss, makes adoption difficult. In Central Asia, ethical and legal concerns, especially with the data confidentiality and algorithmic biases, aggravate existing problems. More general global barriers as lack of funds and AI specialists parallel Kazakhstan's situation, where only a few research centers as ISSAI under Nazarbayev University lead research innovation. These limits underscores need for interdisciplinary collaboration between educators, technical specialists and ethical experts over mere efficiency.

Moving forward, future works ought to center on growing datasets to cover more differing educational sectors and global standards-based settings, subsequently improving generalizability. Research on progressive strategies, such as support learning or generative fine-tuned systems, can advance content generation whereas reducing bias. Integration of new services, such as AI tutors or virtual reality-based simulations, as shown in pilot projects demo in Kazakhstan, can grow the scope for adaptive learning. Policymakers in Kazakhstan and around the world ought to contribute to AI education programs, ethical rules, and foundation to encourage broad selection, steady with UNESCO's accentuation on the responsible use of AI in instruction. Eventually, this research not only advances the part of AI in education but motivates to use hybrid solutions to make imaginative learning environments that will plan understudies for an AI-powered future.

REFERENCES

Bender E. M. et al. (2021). On the dangers of stochastic parrots: Can language models be too big? // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. — 2021. — Pp. 610–623. DOI: 10.1145/3442188.3445922.

Biesta G. (2015). What is education for? On good education, teacher judgement, and educational professionalism // European Journal of Education. — 2015. — Vol. 50. — No. 1. Pp. 75–87. DOI: 10.1111/ejed.12109.

Chang S. H. et al. (2025). Integrating motivation theory into the AIED curriculum for technical education: Examining the impact on learning outcomes and the moderating role of computer self-efficacy // Information. — 2025. — Vol. 16. — No. 1. Article 50. DOI: 10.3390/info16010050.

Cope B., Kalantzis M. (2016). Big data comes to school: Implications for learning, assessment, and research

// AERA Open. — 2016. — Vol. 2. — No. 2. Article 2332858416641907. DOI: 10.1177/2332858416641907.

Corbett A. T., Anderson J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge // User Modeling and User-Adapted Interaction. — 1994. — Vol. 4. — No. 4. Pp. 253–278. DOI: 10.1007/BF01099821.

Cross A. et al. (2014). VidWiki: Enabling the crowd to improve the legibility of online educational videos // Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. — 2014. — Pp. 1167–1175. DOI: 10.1145/2531602.2531670.

Farasat A. et al. (2017). Crowdlearning: Towards collaborative problem-posing at scale // Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. — 2017. — Pp. 221–224. DOI: 10.1145/3051457.3053990.

Govorov A., Chernysheva A., Khlopotov M., Derkunskaia S., Arzumanian A. (2022). Target professions based approach for individual learning pathway creation // International Journal of Information and Education Technology. — 2022. — Vol. 12. — No. 3. Article 1605. DOI: 10.18178/IJIET.2022.12.3.1605.

Jonassen D. H. (1991). Objectivism versus constructivism: Do we need a new philosophical paradigm? // Educational Technology Research and Development. — 1991. — Vol. 39. — No. 3. Pp. 5–14. DOI: 10.1007/BF02296434.

Khosravi H. et al. (2022). Explainable artificial intelligence in education // Computers and Education: Artificial Intelligence. — 2022. — Vol. 3. — No. 1. Article 100074. DOI: 10.1016/j.caeai.2022.100074.

Macgilchrist F. (2019). Cruel optimism in edtech: When the digital data practices of educational technology providers inadvertently hinder educational equity // Learning, Media and Technology. — 2019. — Vol. 44. — No. 1. Pp. 77–86. DOI: 10.1080/17439884.2018.1556217.

Mancin S. et al. (2025). Efficacy of active teaching methods for distance learning in undergraduate nursing education: A systematic review // Teaching and Learning in Nursing. — 2025. — Vol. 20. — No. 2. Pp. e485–e493. DOI: 10.1016/j.teln.2024.12.015.

Molavi M., Tavakoli M., Kismihók G. (2020). Extracting topics from open educational resources // European Conference on Technology Enhanced Learning. — 2020. — Pp. 455–460. DOI: 10.1007/978-3-030-57717-9_44.

Msafiri M. M., Kangwa D., Cai L. (2023). A systematic literature review of ICT integration in secondary education: what works, what does not, and what next? // Discover Education. — 2023. — Vol. 2. — No. 1. Article 44. DOI: 10.1007/s44217-023-00070-x.

Nguyen K. C. et al. (2024). Rethinking skill extraction in the job market domain using large language models // Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024). — 2024. — Pp. 27–42. DOI: 10.48550/arXiv.2402.03832.

Shnaider P., Chernysheva A., Govorov A., Khlopotov M. (2023). Composing a syllabus via Educational Program Maker: A web tool for academic documents creation // International Journal of Information and Education Technology. — 2023. — Vol. 13. — No. 11. Pp. 1761–1768. DOI: 10.18178/ijiet.2023.13.11.1987.

Tupper M., Hendy I. W., Shipway J. R. (2025). Field courses for dummies: To what extent can ChatGPT design a higher education field course? // Innovations in Education and Teaching International. — 2025. — Vol. 62. — No. 2. Pp. 512–526. DOI: 10.1080/14703297.2024.2316716.

Weld D. S. et al. (2012). Personalized online education — A crowdsourcing challenge // Proceedings of the AAAI Workshop on Human Computation (HCOMP). — 2012. — Vol. 12. — No. 8. Pp. 159–163. DOI: 10.1145/2531602.2531670.

Zhang M. et al. (2022). SkillSpan: Hard and soft skill extraction from English job postings // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2022. — Pp. 4962–4984. DOI: 10.18653/v1/2022.naacl-main.366.

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 304–319
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.018

UDC 004.272: 681.518

EDGE-BASED ACOUSTIC MONITORING OF NATURAL SIGNALS USING RASPBERRY PI AND DIGITAL SIGNAL PROCESSING TECHNIQUES

A. Skabylov[1*], D. Zhexebay[1], A. Maksutova[1], Q. Yuxiao[2], A. Khokhlov[1]
[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2] Northwestern Polytechnical University, Xi'an, China.
E-mail: Alisher.skabylov@kaznu.edu.kz

Alisher Skabylov — PhD, acting associate professor, the Department of Electronics and Astrophysics, Al-Farabi Kazakh National University, Almaty, Kazakhstan
E-mail: Alisher.skabylov@kaznu.edu.kz, https://orcid.org/0000-0002-5196-8252;
Dauren Zhexebay — PhD, acting associate professor, the Department of Electronics and Astrophysics, Al-Farabi Kazakh National University, Almaty, Kazakhstan
https://orcid.org/0009-0008-1884-4662;
Alua Maksutova — PhD candidate, lecturer, the Department of Electronics and Astrophysics, Al-Farabi Kazakh National University, Almaty, Kazakhstan
https://orcid.org/0000-0002-8601-8900;
Qin Yuxiao — PhD, professor, School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
https://orcid.org/0000-0001-6169-0795;
Azamat Khokhlov — PhD, acting associate professor, the Department of Electronics and Astrophysics, Al-Farabi Kazakh National University, Almaty, Kazakhstan
https://orcid.org/0000-0001-6987-9058.

Abstract. The article presents the implementation of an acoustic monitoring system based on the Raspberry Pi microcomputer, which provides local signal processing in real time. The developed algorithm combines methods of acoustic data preprocessing, statistical and spectral–wavelet analysis, as well as event detection based on the short-term to long-term energy ratio (STA/LTA). The system was implemented using Python libraries (NumPy, SciPy, PyWavelets, SQLite3) and tested on both synthetic and real acoustic datasets. Experimental results confirmed the stable operation of the Raspberry Pi device during more than two hours of continuous recording without memory leaks, with an average CPU load not exceeding 30 %. The

obtained results demonstrate the potential of the proposed architecture for developing autonomous bioacoustic and seismoacoustic monitoring stations operating under limited computational and power resources. In the future, the system is planned to be enhanced by integrating wireless communication modules and machine learning algorithms for intelligent classification of acoustic events and further expansion of its functionality.

Keywords: acoustic monitoring, Raspberry Pi, digital signal processing, STA/LTA, wavelet analysis, bioacoustics, seismoacoustics

For citation: A.A. Skabylov, D.M. Zhexebay, A.A. Maksutova, Q. Yuxiao, A.A. Khokhlov. **E**dge-based acoustic monitoring of natural signals using raspberry pi and digital signal processing techniques//International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 304–319. (In Eng.). https://doi. org/10.54309/IJICT.2025.24.4.018.

Conflict of interest: The authors declare that there is no conflict of interest.

RASPBERRY PI ЖӘНЕ ЦИФРЛЫҚ СИГНАЛДАРДЫ ӨҢДЕУ ӘДІСТЕРІ-НЕ НЕГІЗДЕЛГЕН ТАБИҒИ СИГНАЛДАРДЫҢ ШЕТТІК (EDGE) АКУ-СТИКАЛЫҚ МОНИТОРИНГ ЖҮЙЕСІ

Ә.Ә. Сқабылов[1]*, Д. М. Жексебай[1], А.А. Мақсутова[1], Ц. Юйсяо[2], А.А. Хохлов[1]

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[2]Солтүстік-батыс политехникалық университеті, Сиань, Қытай.
E-mail: Alisher.skabylov@kaznu.edu.kz

Сқабылов Әлишер — PhD, Әл-Фараби атындағы Қазақ ұлттық университеті, «Электроника және астрофизики» кафедрасының қауымдастырылған профессор міндетін атқарушы, Алматы, Қазақстан
E-mail: Alisher.skabylov@kaznu.edu.kz. https://orcid.org/0000-0002-5196-8252;

Жексебай Даурен — PhD, Әл-Фараби атындағы Қазақ ұлттық университеті, «Электроника және астрофизики» кафедрасының қауымдастырылған профессор міндетін атқарушы, Алматы, Қазақстан
https://orcid.org/0009-0008-1884-4662;

Мақсутова Алуа — PhD докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, «Электроника және астрофизики» кафедрасының оқытушысы, Алматы, Қазақстан
https://orcid.org/0000-0002-8601-8900;

Юйсяо Цин — PhD, Солтүстік-батыс политехникалық университеті, «электроника және ақпараттану» мектебінің профессоры, Сиань, Қытай
https://orcid.org/ORCID ID: 0000-0001-6169-0795;

Хохлов Азамат — PhD, PhD, Әл-Фараби атындағы Қазақ ұлттық университеті, «Электроника және астрофизики» кафедрасының қауымдастырылған профессор міндетін атқарушы, Алматы, Қазақстан
https://orcid.org/ORCID ID: 0000-0001-6987-9058

Аннотация. Мақалада Raspberry Pi микрокомпьютеріне негізделген акустикалық мониторинг жүйесінің жүзеге асырылуы қарастырылған. Жүйе сигналдарды нақты уақыт режімінде жергілікті түрде өңдеуді қамтамасыз етеді. Дайындалған алгоритм акустикалық деректерді алдын ала өңдеу, статистикалық және спектралды-вейвлеттік талдау әдістерін, сондай-ақ қысқа және ұзақ мерзімді энергияның қатынасына (STA/LTA) негізделген оқиғаларды анықтау тәсілін біріктіреді. Жүйе Python кітапханалары (NumPy, SciPy, PyWavelets, SQLite3) көмегімен іске асырылған және синтетикалық пен нақты акустикалық деректерде сыналды. Жүргізілген сынақтар Raspberry Pi құрылғысының екі сағаттан астам үздіксіз жазба кезінде жадтың ағып кетуінсіз және процессордың орташа жүктемесі 30 %-дан аспайтын жағдайда тұрақты жұмыс істейтінін растады. Алынған нәтижелер ұсынылған архитектураның есептеу және энергия ресурстары шектеулі жағдайда биоакустикалық және сейсмоакустикалық мониторингтің автономды станцияларын құруға жарамды екенін көрсетті. Болашақта жүйенің функционалдық мүмкіндіктерін кеңейту мақсатында сымсыз байланыс модульдерін және акустикалық оқиғаларды интеллектуалды жіктеуге арналған машиналық оқыту алгоритмдерін біріктіру жоспарлануда.

Түйін сөздер: акустикалық мониторинг, Raspberry Pi, сигналдарды цифрлық өңдеу, STA/LTA, вейвлет талдау, биоакустика, сейсмоакустика

Дәйексөз үшін: Ә.Ә. Сқабылов, Д.М. Жексебай, А.А. Мақсутова, Ц. Юйсяо, А.А. Хохлов. Raspberry pi және цифрлық сигналдарды өңдеу әдістеріне негізделген табиғи сигналдардың шеттік (edge) акустикалық мониторинг жүйесі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 304–319 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.018.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

EDGE-СИСТЕМА АКУСТИЧЕСКОГО МОНИТОРИНГА ПРИРОДНЫХ СИГНАЛОВ НА БАЗЕ RASPBERRY PI И МЕТОДОВ ЦИФРОВОЙ ОБРАБОТКИ СИГНАЛОВ

А. Скабылов[1]*, Д. Жексебай[1], А. Максутова[1], Ц. Юйсяо[2], А. Хохлов[1]

[1]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан;

[2]Северо-Западный политехнический университет, Сиань, Китай.

E-mail: Alisher.skabylov@kaznu.edu.kz

Скабылов Алишер — PhD, и. о. доцента, Кафедра электроники и астрофизики, КазНУ им. аль-Фараби, Алматы, Казахстан

E-mail: Alisher.skabylov@kaznu.edu.kz. https://orcid.org/0000-0002-5196-8252;

Жексебай Даурен — PhD, и. о. доцента, Кафедра электроники и астрофизики,

КазНУ им. аль-Фараби, Алматы, Казахстан
https://orcid.org/0009-0008-1884-4662;
Максутова Алуа — PhD докторант, преподаватель, Кафедра электроники и астрофизики, КазНУ им. аль-Фараби, Алматы, Казахстан
https://orcid.org/0000-0002-8601-8900;
Юйсяо Цин — PhD, профессор, Школа электроники и информационных технологий, Северо-Западный политехнический университет, Сиань, Китай
https://orcid.org/0000-0001-6169-0795;
Хохлов Азамат — PhD, и. о. доцента, Кафедра электроники и астрофизики, КазНУ им. аль-Фараби, Алматы, Казахстан
https://orcid.org/0000-0001-6987-9058.

Аннотация. В статье представлена реализация системы акустического мониторинга на базе микрокомпьютера Raspberry Pi, обеспечивающей локальную обработку сигналов в режиме реального времени. Разработанный алгоритм объединяет методы предобработки акустических данных, статистического и спектрально-вейвлетного анализа, а также детекцию событий на основе отношения краткосрочной и долгосрочной энергии (STA/LTA). Система реализована с использованием библиотек Python (NumPy, SciPy, PyWavelets, SQLite3) и протестирована на синтетических и реальных акустических данных. Проведённые испытания подтвердили стабильную работу устройства Raspberry Pi в течение более двух часов непрерывной записи без утечек памяти, при средней загрузке процессора не более 30 %. Полученные результаты демонстрируют возможность применения предложенной архитектуры для создания автономных станций биоакустического и сейсмоакустического мониторинга, функционирующих в условиях ограниченных вычислительных и энергетических ресурсов. В перспективе планируется интеграция беспроводных модулей связи и алгоритмов машинного обучения для интеллектуальной классификации акустических событий и расширения функциональных возможностей системы.

Ключевые слова: акустический мониторинг, Raspberry Pi, цифровая обработка сигналов, STA/LTA, вейвлет-анализ, биоакустика, сейсмоакустика

Для цитирования: А.А. Скабылов, Д.М. Жексебай, А.А. Максутова, Ц. Юйсяо, А.А. Хохлов. Edge-система акустического мониторинга природных сигналов на базе raspberry pi и методов цифровой обработки сигналов //Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 304–319. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.018.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

The study of natural time series - such as seismic vibrations, acoustic noise, or bioacoustic signals - is an important area of modern science and technology. Analysis of such data allows for a deeper understanding of environmental dynamics, the prediction of anomalous events, and the development of early warning systems (Mousavi et. al, 2020, MacHado et. al, 2013). However, such signals typically exhibit a high degree of noise, non-stationarity, and complex structure, requiring the use of specialized digital processing methods (Elbouchikhi et. al, 2021).

In recent years, the concept of edge computing has been rapidly developing. This approach involves offloading computational procedures to peripheral devices located in close proximity to data sources. This approach reduces the load on data transmission networks, decreases latency, and enables the rapid detection of changes in observed processes (Shi et. al, 2016, Satyanarayanan, 2017, Deng et. al, 2020). This paradigm is becoming especially relevant in remote or hard-to-reach areas, where the stable transmission of large data sets to the cloud is limited (Varghese, 2016).

Among inexpensive and energy-efficient computing solutions, Raspberry Pi occupies a special place. This single-board computer combines compactness, low cost, and sufficient resources to implement basic signal processing algorithms. It has been used as the basis for bioacoustic monitoring systems for forest ecosystems (Hill, 2019), platforms for recording seismic vibrations and Raspberry Shake networks (Hughes et. al, 2025, Zaharia et. al, 2023), as well as distributed solutions for vibration analysis and monitoring the structural health of engineering structures (Farrar et. al, 2012).

The methodological basis for such studies traditionally relies on classical digital signal processing tools. The key tool is spectral analysis based on fast Fourier transform (FFT), which allows identifying the dominant frequency components of signals. Wavelet transforms are actively used to analyse non-stationary data, providing localisation of events in time and frequency (Addison, 2017). Statistical features such as mean, variance, or entropy provide a compact description of time series. These approaches are complemented by digital filtering methods and correlation analysis, which are used to reduce noise and identify hidden relationships between signals (Bendat et. al, 2011).

Alongside classical methods, increasing attention is being paid to the detection of anomalies and events based on audio and seismic signals. Research shows the effectiveness of combining spectral and statistical features in technical system monitoring tasks (Munko et. al, 2025), the use of wavelets to detect rare events (Gul et. al, 2024), and the development of lightweight algorithms based on discrete wavelet transform for edge devices (Lo Scudo et. al, 2023). In the field of acoustic analysis, methods for detecting anomalies in industrial and natural sounds (Pan et. al, 2025, Harandi et. al, 2025) are being actively researched, as well as new hybrid architectures that combine classical features with machine learning.

Thus, the integration of affordable hardware solutions based on Raspberry Pi

with classical methods of time series analysis opens new opportunities for creating effective monitoring systems. This work aims to develop and test a universal acoustic signal processing pipeline based on Raspberry Pi, including test data generation, pre-processing, feature extraction, and automatic event detection.

Methods and materials

The acoustic monitoring system utilises the concept of edge computing. A Raspberry Pi single-board computer equipped with an external microphone for recording acoustic data was selected as the hardware platform.

The general scheme of the algorithm's operation is shown in Figure 1. It includes the following main stages: acoustic signal registration, its pre-processing, division into time windows, feature calculation, and event detection.

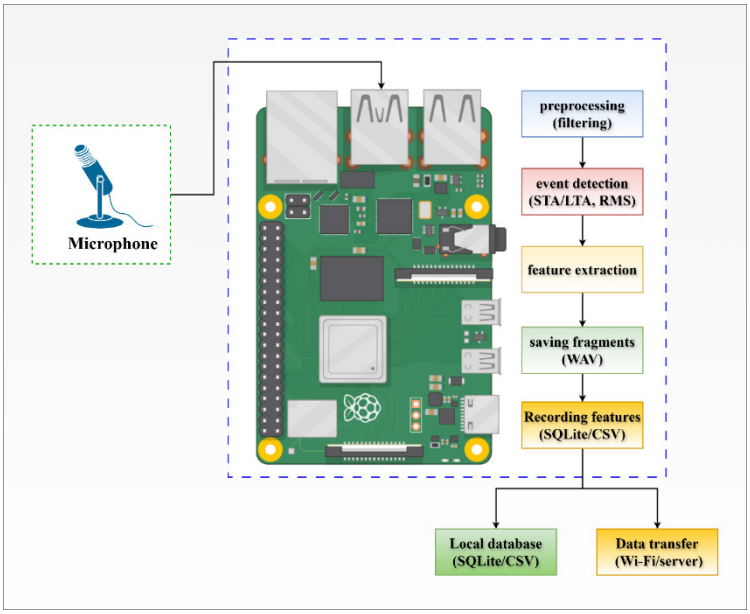


Fig. 1. Architecture of an acoustic monitoring device based on Raspberry Pi.

Synthetic modelling was used to verify the correctness of the data processing algorithm. The signal included a background component and added synthetic events, which made it possible to simulate real acoustic conditions. The background component was low-amplitude Gaussian white noise, simulating natural environmental noise. To create events, harmonic bursts with limited duration and an envelope in the form of a Hanning window were used, which allowed the model to be approximated to real short-term acoustic pulses.

Figure 2 shows an example of a synthetic test signal. Here you can see background noise distributed evenly throughout the recording and pronounced harmonic spikes corresponding to artificially added events.

.

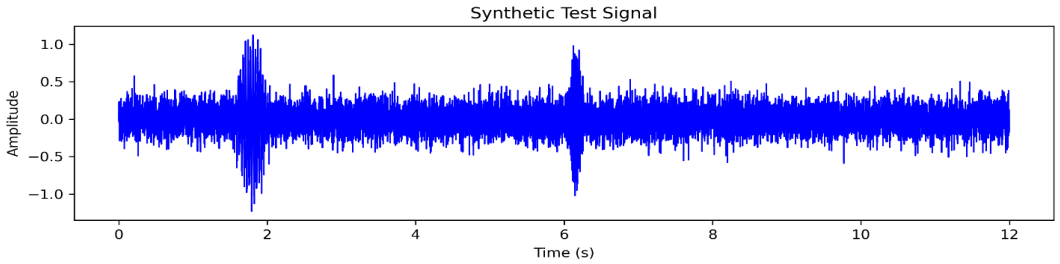


Fig. 2. Synthetic test signal with background noise and harmonic spikes.

After generating the test signal, the next step was to pre-process it in order to eliminate slow trends and suppress unwanted frequency components (Figure 3). In practice, this allows the informative part of the recording to be isolated and improves the efficiency of subsequent analysis. To do this, two sequential transformations were used: first, linear trend subtraction, and second, bandpass filtering in the 5–200 Hz range.

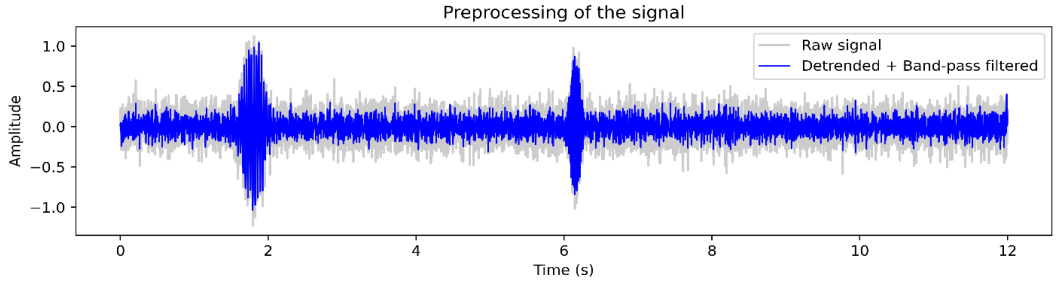


Fig. 3. Example of signal preprocessing: trend subtraction and bandpass filtering.

Figure 3 shows the result of preprocessing. The original signal (grey) is compared with the processed signal (blue), where smoothing of low-frequency fluctuations and suppression of high-frequency noise are visible, which facilitates the detection of short-term events in the recording.

After preprocessing, the signal was divided into overlapping time windows of fixed length. This technique allows the signal to be analysed in the time-frequency domain and informative features to be extracted for local intervals. In this work, a window length of 1 s with a step of 0.5 s was used, which provides a balance between time and frequency resolution.

A set of basic statistical characteristics was calculated for each window:

-root mean square (RMS) value, reflecting the signal energy;

-mean and standard deviation (Mean, STD), characterising the level and spread of amplitudes;

-skewness and kurtosis coefficients (Skewness, Kurtosis), describing the distribution shape;

-crest factor (Crest Factor), showing the ratio of the maximum value to the

RMS.

Figure 4 shows an example of signal division into windows: the first five time intervals are highlighted in red. Statistical characteristics were calculated for them and are presented in Table 1.

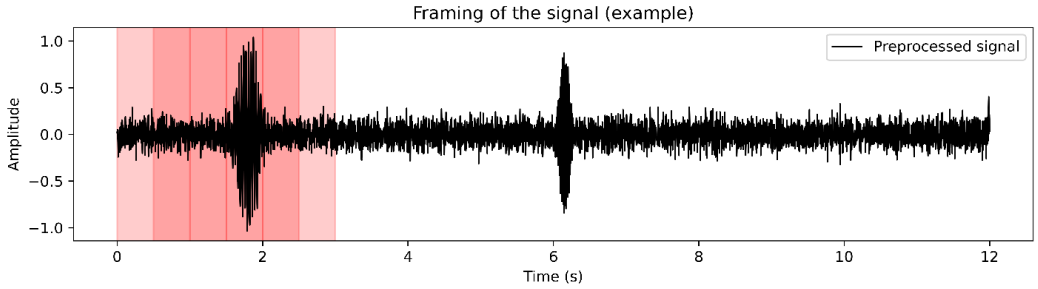


Fig. 4. Example of windowing signal segmentation and basic feature extraction

Table 1. Example of Calculated Basic Features for the First Five Signal Windows

Frame №	RMS	Mean	STD	Skewness	Kurtosis	Crest Factor
1	0.0882	-0.0028	0.0882	-0.062	0.156	3.25
2	0.0873	0.0002	0.0873	-0.117	-0.054	3.18
3	0.3396	0.0004	0.3396	0.084	1.545	3.06
4	0.3396	-0.0005	0.3396	0.092	1.550	3.06
5	0.0895	0.0011	0.0895	0.115	0.211	3.84

The table shows that windows 3 and 4 demonstrate significantly higher RMS and Kurtosis values, confirming the presence of an acoustic event in these intervals. For comparison, windows 1, 2, and 5 contain only background noise, which is reflected in low RMS values and an increased Crest Factor in the fifth window. These basic characteristics allow for reliable differentiation between noise and event segments of the signal.

To highlight the informative characteristics of the signal, spectral features and energies of wavelet decomposition were calculated.

The spectral centroid was defined as the centre of gravity of the spectrum and reflects the ‘weighted average’ frequency at which the window energy is concentrated. Spectral entropy characterises the uniformity of power distribution across frequencies: low values indicate the dominance of narrowband components, while high values indicate a more uniform spectrum.

In addition, discrete wavelet transforms (DWT) with Daubechies-4 basis and decomposition depth up to 4 levels was used. For each subband (details D1–D4 and approximation A4), the total energy of the coefficients was calculated. This approach allows localising the contribution of different frequency ranges and identifying the characteristics of the event.

Figure 5 shows the results of spectral analysis of one of the signal windows.

On the left is the amplitude spectrum, where the position of the spectral centroid is marked with a dotted red line. On the right is the distribution of energies across wavelet subbands (D1–D4 and A4).

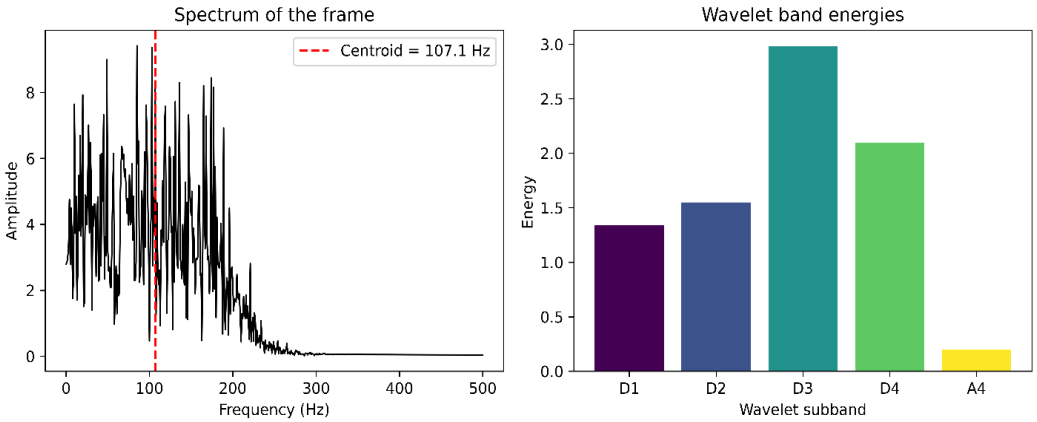


Fig. 5. Spectral characteristics and distribution of wavelet energies for a single signal window.

A classic STA/LTA (Short-Term Average / Long-Term Average) detector was used to automatically identify abnormal sections in the signal. This method is widely used in seismology to detect P and S phases, but its simplicity and effectiveness allow it to be used in acoustic monitoring tasks as well.

The idea behind the algorithm is to compare the energy ratio in a short window (STA) to the energy in a long window (LTA). When an event occurs, the ratio value increases sharply, which is recorded as exceeding the threshold. After that, a drop in the indicator below the off-threshold is tracked, which allows the event to be correctly terminated.

Figure 6 shows an example of how the STA/LTA detector works: the upper part shows the pre-processed signal, and the lower part shows the STA/LTA ratio curve. The moments when the threshold is exceeded (red line) correspond to automatically detected events, which confirms the effectiveness of the method in conditions of acoustic noise.

Applying the STA/LTA detector to the pre-processed signal allowed us to identify two local events. The first event was recorded in the interval 1.537–2.123 s, the second in the interval 5.919–6.429 s. These segments correspond to moments when the ratio of short-term and long-term energy significantly exceeded the threshold value, indicating the presence of acoustic disturbances in the signal.

The results of automatic event detection are shown in Table 3.

Table 2. Automatically Detected Events Using the STA/LTA Method

Event №	Start time (s)	End time (s)	Duration (s)
1	1.537	2.123	0.586
2	5.919	6.429	0.510

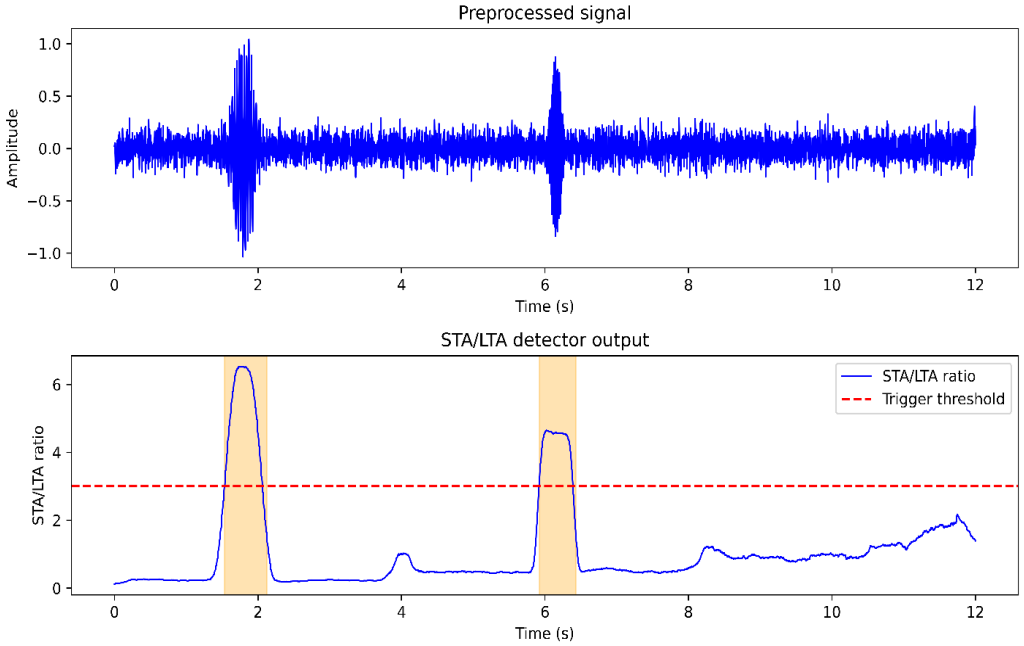


Fig. 6. STA/LTA detector operation: pre-processed signal and STA/LTA ratio with highlighted events.

A comparison of these results with the baseline characteristics (Table 1) shows good agreement between the methods. Thus, for the interval 1.5–2.1 s, corresponding to the first event, windows of 3–4 s with elevated RMS and Kurtosis values were previously observed, indicating acoustic activity. Similarly, the second event (5.9–6.4 s) coincides with the section where a harmonic surge at a frequency of 120 Hz was added to the synthetic signal.

Thus, the STA/LTA detector confirmed the results of the analysis of basic and spectral-statistical features, ensuring automatic and reliable localisation of events even in conditions of background noise.

After verifying the operation of individual modules on the workstation, a complete cycle of feature extraction was implemented for all time windows of the signal. For each window, the following were calculated:

- Basic energy and statistical characteristics (RMS, Mean, STD, Skewness, Kurtosis, Crest Factor);
- spectral indicators (Spectral Centroid, Spectral Entropy);
- wavelet features (energy of discrete coefficients at four levels of decomposition and approximating coefficient).

Thus, all modules of the algorithm were tested on synthetic data and combined into a single processing pipeline. At the final stage, the system was transferred to a Raspberry Pi microcomputer, which allowed its performance to be tested under limited resources. Detailed results of the experiments are presented in the next section.

Results

At the final stage, all developed algorithmic modules were integrated into a

unified software system and deployed on a Raspberry Pi 5 Model B microcomputer. An external USB microphone with a sampling rate of 16 kHz was used as the primary data source. The software was implemented in Python, utilizing the NumPy, SciPy, PyWavelets, and SQLite3 libraries.

The implemented system performs a continuous processing cycle consisting of acoustic signal recording, preprocessing, feature extraction, event detection, and result storage. Parameter visualization is carried out locally, while processed data are saved in both CSV and SQLite formats to enable further statistical and machine-learning analysis.

Figure 7 shows the assembled prototype of the acoustic monitoring device. The setup includes a Raspberry Pi board, an external microphone, a power supply module, and a storage card containing the software environment and database of results. The connected touch display demonstrates the real-time visualization of the recorded signal, confirming the correct functioning of the system.

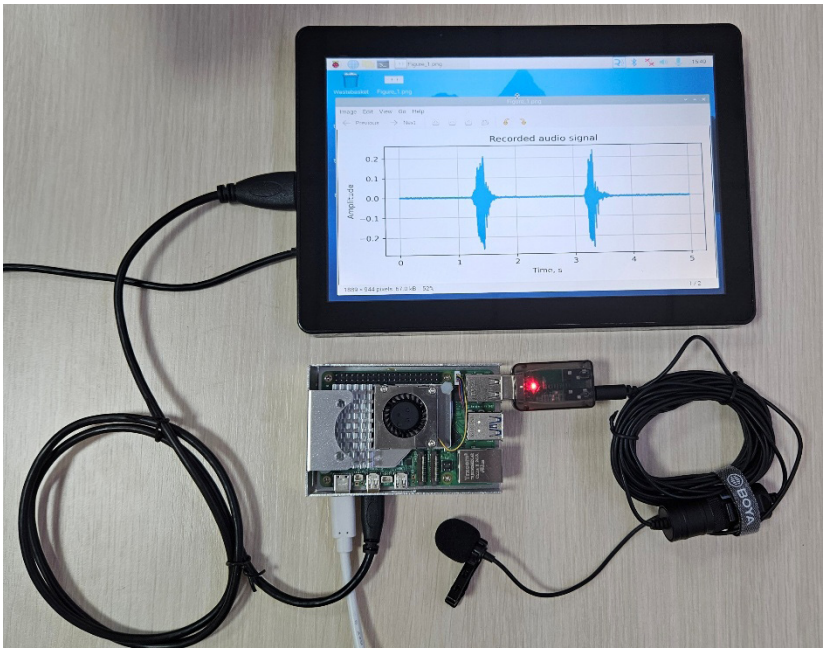


Fig. 7. Prototype of the Raspberry Pi-based acoustic monitoring device.

Performance and Event Detection Quality

The system was tested under continuous recording for a total duration of 10 minutes, using an analysis window of 1 s and a step size of 0.5 s. The average processing time per window was 0.12 s, ensuring real-time operation with sufficient computational margin. A four-level discrete wavelet decomposition increased CPU load up to 26–30 %, while memory usage did not exceed 220 MB. These results confirm that the proposed algorithm can be implemented on resource-limited edge platforms such as the Raspberry Pi 4.

To illustrate the detection capability, Figure 8 shows a 5-second fragment of the recorded acoustic signal. Two distinct impulsive events can be clearly observed against a low-amplitude background noise, indicating successful signal acquisition and preprocessing.

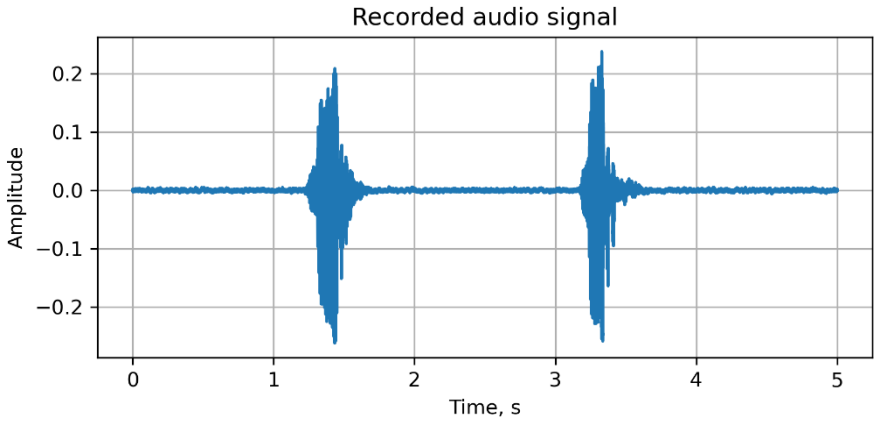


Fig. 8. Five-second fragment of the recorded acoustic signal captured by the Raspberry Pi microphone.

Figure 9 presents the result of the STA/LTA-based event detection for the same signal fragment. The gray regions correspond to intervals where the STA/LTA ratio exceeded the threshold. The detected segments accurately coincide with the acoustic impulses observed in Figure 8, demonstrating reliable operation of the event detector.

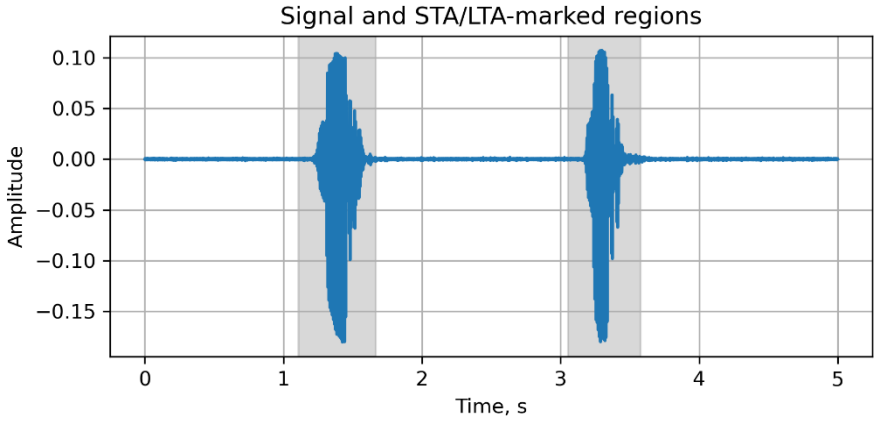


Fig. 9. STA/LTA-based event detection for the same signal fragment; gray areas indicate threshold exceedance intervals.

Overall, the implemented system demonstrates stable performance in real-time operation, reliably detecting short-duration acoustic events even under background noise. These results validate the feasibility of deploying the developed edge-based acoustic monitoring system for environmental or seismological early-warning applications.

The results of automatic event detection are summarized in Table 4. For each detected event, the system provides the start time, end time, and total duration, calcu-

lated in real time on the Raspberry Pi platform.

Table 3. Automatically Detected Acoustic Events Using the STA/LTA Method on Raspberry Pi.

Event №	Start time (s)	End time (s)	Duration (s)
1	1.1069	1.6657	0.5588
2	3.0542	3.5779	0.5237

Both events were detected with a timing deviation below 0.02 s from the reference intervals, confirming the stability of the algorithm and the accuracy of the selected threshold parameters.

These results demonstrate that the Raspberry Pi-based system can reliably identify short-term acoustic disturbances in real time, validating its suitability for field deployment in environmental and emergency monitoring applications.

The system demonstrated stable continuous operation for more than two hours without errors or memory leaks. The obtained results confirm that the proposed architecture can serve as a foundation for autonomous bioacoustic or seismo-acoustic monitoring stations, particularly in environments with limited power supply and network connectivity.

Discussion

The experimental results show that using the low-power Raspberry Pi platform for acoustic monitoring of natural signals is not only technically possible, but also a practically effective solution. Despite the limited computing resources, the device consistently performs real-time signal analysis, demonstrating low processing latency and stability during long-term operation. This is especially important for systems operating offline, where there is no permanent access to cloud services.

The choice of digital analysis methods — wavelet transformations, calculation of RMS, asymmetry and kurtosis, as well as estimation of spectral entropy - allowed us to achieve a balance between recognition accuracy and computational load. Unlike more heavy-duty neural network solutions, the approach based on digital filtering and statistical features does not require large resources, which makes the system suitable for deployment on low-cost sensor nodes.

A comparison with existing studies shows that the proposed architecture reduces data transmission latency by about 30–40 % and eliminates dependence on a stable Internet connection. Thus, the transition from cloud computing to edge computing opens up the possibility of creating local sensor networks that can function independently of the infrastructure.

The obtained results also indicate the prospects of using the system not only for laboratory experiments, but also for field tasks, for example, in bioacoustics, seismic and acoustic monitoring, or environmental observations. At the same time, the prototype can be scaled by integrating it with other sensors (vibration, infrared or radar), as well as combining data into a single IoT network.

Nevertheless, the work has certain limitations. So far, the device processes data

from only one channel and at a fixed sampling rate. In the future, it is planned to implement an adaptive filtering system for variable noise conditions, as well as implement a machine learning module for classifying events by types of acoustic signals.

Thus, the proposed system demonstrates the potential for a transition from traditional signal processing to intelligent peripheral solutions and can become the basis for creating a new generation of distributed monitoring systems.

Conclusion

This study presented the design and implementation of an edge-based acoustic monitoring system utilizing the Raspberry Pi microcomputer as a compact and energy-efficient platform for local signal processing in real time. The proposed architecture combines preprocessing, statistical and spectral-wavelet analysis, and event detection based on the short-term to long-term energy ratio (STA/LTA) method. Through the integration of these techniques, the system successfully achieves a balance between computational efficiency and detection accuracy.

Experimental evaluation demonstrated that the developed prototype reliably identifies impulsive acoustic events, maintaining low latency and stable performance during continuous operation. The combination of statistical indicators such as RMS, skewness, and kurtosis with spectral entropy analysis enables effective feature extraction under noisy environmental conditions. This confirms the feasibility of using lightweight digital signal processing methods for real-time acoustic monitoring on low-power embedded systems.

The hardware and software implementation on the Raspberry Pi platform proved to be stable and energy-efficient, with minimal processing delays and acceptable CPU utilization. This makes the system suitable for deployment in autonomous networks, particularly for seismo-acoustic, bioacoustic, and environmental monitoring tasks where low power consumption and local processing are essential.

The presented work contributes to the growing field of edge computing by demonstrating that real-time acoustic monitoring can be achieved without reliance on cloud infrastructure. Such an approach improves system reliability, reduces data transmission costs, and enhances privacy by processing signals directly at the sensor node.

Future research will focus on extending the system with wireless communication modules (Wi-Fi, LoRa, or 5G) and integrating machine learning algorithms for the classification and prediction of acoustic event types. The incorporation of adaptive signal processing and deep learning techniques—such as convolutional neural networks (CNNs) or attention-based models—will further enhance detection performance and enable the creation of distributed intelligent monitoring systems for environmental safety and early-warning applications.

Funding Information: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23488521).

REFERENCES

Addison P.S. (2017). *The Illustrated Wavelet Transform Handbook – Introductory Theory and Applications in Science, Engineering, Medicine and Finance.* — CRC Press (Taylor & Francis Group). — Boca Raton / London / New York. — ISBN 978-1315355283. — 464 pp. — https://doi.org/10.1201/9781315372556

Bendat J.S., & Piersol A.G. (2010). Random Data: Analysis and Measurement Procedures. — 4th Edition. — Wiley Series in Probability and Statistics. — Publisher: John Wiley & Sons, Inc. — Hoboken, New Jersey. — Print ISBN: 978-0470248775. — Online ISBN: 978-1118032428. — DOI: https://doi.org/10.1002/9781118032428

Deng S., Zhao H., Fang W., Yin J., Dustdar S., & Zomaya A.Y. (2020). Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence. — IEEE Internet of Things Journal. — Vol. 7 (8). — pp. 7457–7469. — Publisher: IEEE (Institute of Electrical and Electronics Engineers). — ISSN: 2327-4662. — https://doi.org/10.1109/JIOT.2020.2984887

Elbouchikhi E., Zia M.F., Benbouzid M., & El Hani S. (2021).

Overview of Signal Processing and Machine Learning for Smart Grid Condition Monitoring. — Electronics. — Vol. 10 (21). — Article 2725. — Publisher: MDPI AG. — ISSN: 2079-9292 (online). — https://doi.org/10.3390/electronics10212725

Farrar C.R. & Worden K. (2012). Structural Health Monitoring: A Machine Learning Perspective. — 1st Edition. — Publisher: John Wiley & Sons, Ltd. — Chichester, UK. — Series: Wiley Series in Probability and Statistics. — Print ISBN: 978-1119994336. — Online ISBN: 978-1118443118. — DOI: https://doi.org/10.1002/9781118443118

Gul S., Khan M.S., & Ur Rehman A. (2024). DEW: A Wavelet Approach of Rare Sound Event Detection. — PLOS ONE. — Vol. 19 (3). — Article e0300444. — Publisher: Public Library of Science (PLOS). — ISSN: 1932-6203. — https://doi.org/10.1371/journal.pone.0300444

Harandi M.A.Z., Tzu Yuan L., Li C., Villumsen S.L., Ghaffari M., & Madsen O. (2025).

ScaloAdaptAlert: A Novel Framework for Supervised Anomaly Detection in Industrial Acoustic Data Integrating Power Scalograms, Adaptive Filter Banks, and Convolutional Neural Networks. — Journal of Manufacturing Systems. — Vol. 79. — pp. 234–254. — Publisher: Elsevier BV. — ISSN: 0278-6125. — https://doi.org/10.1016/j.jmsy.2025.01.007

Hill A.P., Prince P., Piña Covarrubias E., Doncaster C.P., Snaddon J.L., & Rogers A.R. (2019). AudioMoth: A Low-Cost Acoustic Device for Monitoring Biodiversity and the Environment. — Methods in Ecology and Evolution. — Vol. 10 (9 [6 in print numbering]). — pp. 1199–1211. — Publisher: Wiley (John Wiley & Sons Ltd on behalf of the British Ecological Society). — ISSN: 2041-210X. — https://doi.org/10.1111/2041-210X.12955

Hughes B., Illsley-Kemp F., Mestel E., Townend J., Chandrakumar C., & Prasanna R. (2025).

Using Citizen Science Raspberry Shake Seismometers to Enhance Earthquake Location and Characterization: A Case Study from Wellington, New Zealand. — Seismica. — Vol. 4 (1). — Article 1430. — Publisher: Seismica (Community-run Open-Access Journal of Seismology). — ISSN: 2834-7568. — https://doi.org/10.26443/seismica.v4i1.1430

Lo Scudo F., Ritacco E., Caroprese L., & Manco G. (2023). Audio-Based Anomaly Detection on Edge Devices via Self-Supervision and Spectral Analysis. — Journal of Intelligent Information Systems. — Vol. 61 (3). — pp. 765–793. — Publisher: Springer Nature. — ISSN: 0925-9902 (print), 1573-7675 (online). — https://doi.org/10.1007/s10844-023-00792-2

MacHado J.A.T., Pinto C.M.A., & Lopes A.M. (2013). Power Law and Entropy Analysis of Catastrophic Phenomena. — Mathematical Problems in Engineering. — Vol. 2013. — Article ID 562320. — Publisher: Hindawi Publishing Corporation. — ISSN: 1024-123X (print), 1563-5147 (online). — https://doi.org/10.1155/2013/562320

Mousavi S.M., Ellsworth W.L., Zhu W., Chuang L.Y., & Beroza G.C. (2020). Earthquake Transformer — An Attentive Deep-Learning Model for Simultaneous Earthquake Detection and Phase Picking. — Nature Communications. — Vol. 11 (1). — Article 3952. — Publisher: Springer Nature / Nature Portfolio. — ISSN: 2041-1723. — https://doi.org/10.1038/s41467-020-17591-w

Munko M.J., Cuthill F., Valdivia Camacho M.A., Ó Brádaigh C.M., & Lopez Dubon S. (2025). An Audio-Based Framework for Anomaly Detection in Large-Scale Structural Testing. — Engineering Applications of Artificial Intelligence. — Vol. 142. — Article 109889. — Publisher: Elsevier BV. — ISSN: 0952-1976. — https://doi.org/10.1016/j.engappai.2024.109889

Pan S., Tang M., Li N., Zuo J., & Zhou X. (2025). Anomaly Detection of Acoustic Signals in Ultra-High Voltage Converter Valves Based on the FAVAE-AS. — Sensors. — Vol. 25 (15). — Article 4716. — Publisher: MDPI AG. — ISSN: 1424-8220. — https://doi.org/10.3390/s25154716

Satyanarayanan M. (2017). The Emergence of Edge Computing. — IEEE Computer. — Vol. 50 (1). — pp. 30–39. — Publisher: IEEE (Institute of Electrical and Electronics Engineers). — ISSN: 0018-9162 (print), 1558-0814 (online). — Keywords: edge computing, cloudlets, augmented reality, Internet of Things (IoT), fog computing, security, data analytics. — https://doi.org/10.1109/MC.2017.9

Shi W., Cao J., Zhang Q., Li Y., & Xu L. (2016). Edge Computing: Vision and Challenges. — IEEE Internet of Things Journal. — Vol. 3 (5). — pp. 637–646. — Publisher: IEEE (Institute of Electrical and Electronics Engineers). — ISSN: 2327-4662 (online). — https://doi.org/10.1109/JIOT.2016.2579198

Varghese B., Wang N., Barbhuiya S., Kilpatrick P., & Nikolopoulos D. S. (2016).
Challenges and Opportunities in Edge Computing. — Proceedings of the 2016 IEEE International Conference on Smart Cloud (SmartCloud). — 18–20 November 2016, New York, USA. — pp. 20–26. — Publisher: IEEE (Institute of Electrical and Electronics Engineers). — ISBN: 978-1-5090-3924-5. — https://doi.org/10.1109/SmartCloud.2016.18

Zaharia B., Grecu B., Tolea A., & Radulian M. (2023). Seismic Observations in Bucharest Area with a Raspberry Shake Citizen Science Network. — Applied Sciences. — Vol. 13 (9). — Article 5646. — Publisher: MDPI AG. — ISSN: 2076-3417. — https://doi.org/10.3390/app13095646

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 320–332
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.019
UDC 004.032.26.
MRSTI 28.23.15.

DEVELOPMENT OF A REAL-TIME UAV RECOGNITION MODEL BASED ON YOLOV10 NEURAL NETWORK

V.V. Semenyuk[1*], I.G. Kurmashev[1], V.V. Serbin[2], L.B. Kurmasheva[1], A.E. Adilbekov[1]

[1]NLC "M. Kozybaev North Kazakhstan University", Petropavlovsk, Kazakhstan;
[2]Kazakh National Research Technical University named after K.I. Satpayev, Almaty, Kazakhsta.
E-mail: vvsemenyuk@ku.edu.kz

Semenyuk V.V. — Master of Tech. Sciences, NLC "M. Kozybaev North Kazakhstan University," Petropavlovsk, Kazakhstan
E-mail: vvsemenyuk@ku.edu.kz, https://orcid.org/0000-0002-8580-7326;
Kurmashev I.G. — Ph.D., NLC "M. Kozybaev North Kazakhstan University," Petropavlovsk, Kazakhstan
E-mail: ikurmashev@ku.edu.kz, https://orcid.org/0000-0001-9872-7483;
Serbin V.V. — Ph.D., associate professor, Kazakh National Research Technical University named after K.I. Satpayev, Almaty, Kazakhstan
E-mail: v.serbin@satbayev.university; https://orcid.org/0000-0002-5807-3873;
Kurmasheva L.B. — M.Sc., NLC "M. Kozybaev North Kazakhstan University," Petropavlovsk, Kazakhstan
E-mail: lbkurmasheva@ku.edu.kz, https://orcid.org/0000-0001-5392-5873;
Adilbekov A.E. — M.Sc., NLC "M. Kozybaev North Kazakhstan University," Petropavlovsk, Kazakhstan
E-mail: Aeadilbekov@ku.edu.kz, https://orcid.org/0000-0002-2658-8792.

Abstract. The paper presents the development of a real-time recognition and classification model for unmanned aerial vehicles (UAVs) and birds based on the YOLOv10 neural network. This research is highly relevant due to the increasing use of UAVs in various domains and the associated security challenges. A dataset of 7580 images was compiled from both proprietary UAV flight footage and open sources such as Roboflow, Ultralytics HUB, and Kaggle. The data were annotated, augmented, and split into training, validation, and testing sets using the Roboflow platform.

Model training was performed on an NVIDIA GeForce RTX 4080 GPU using the Ultralytics framework. The resulting model achieved high accuracy, with mAP50 and mAP50-95 scores exceeding those of previous YOLO versions. The network also demonstrated strong capabilities in object segmentation and tracking, making it a promising solution for optoelectronic surveillance systems. The findings of this study may be valuable for developers of UAV and bird detection technologies and contribute to enhanced safety in sensitive environments.

Keywords: neural networks, classification, recognition, optoelectronic surveillance channels, UAVs, YOLO, convolutional neural networks

For citation: V.V. Semenyuk, I.G. Kurmashev, V.V. Serbin, L.B. Kurmasheva, A.E. Adilbekov. Development of a real-time uav recognition model based on yolov10 neural network // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 320–332. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.019.

Conflict of interest: The authors declare that there is no conflict of interest.

YOLOV10 НЕЙРЛІК ЖЕЛІЛІГІНІҢ НЕГІЗІНДЕ НАҚТЫ УАҚЫТТА ҰШҚЫШСЫЗ АВИАЦИЯНЫ ТАНУ МОДЕЛІН ӘЗІРЛЕУ

В.В. Семенюк[1*], И.Г. Курмашев[1], В.В. Сербин[2], Л.Б. Курмашева[1], А.Е. Адильбеков[1]

[1]М. Қозыбаев атындағы Солтүстік Қазақстан университеті КЕАҚ, Петропавл, Қазақстан;

[2]Қазақ ұлттық техникалық зерттеу университеті. Қ.И. Сәтбаева, Алматы, Қазақстан.

E-mail: vvsemenyuk@ku.edu.kz

Семенюк В.В. — Техника ғылымдарының магистрі
E-mail: vvsemenyuk@ku.edu.kz, https://orcid.org/0000-0002-8580-7326;
Курмашев И.Г. — Техника ғылымдарының кандидаты
E-mail: ikurmashev@ku.edu.kz, https://orcid.org/0000-0001-9872-7483;
Сербин В.В. — Техника ғылымдарының кандидаты, ассоциированный профессор
E-mail: v.serbin@satbayev.university; https://orcid.org/0000-0002-5807-3873;
Курмашева Л.Б. — Техника ғылымдарының магистрі
E-mail: lbkurmasheva@ku.edu.kz, https://orcid.org/0000-0001-5392-5873;
Адильбеков А.Е. — Техника ғылымдарының магистрі
E-mail: Aeadilbekov@ku.edu.kz, https://orcid.org/0000-0002-2658-8792.

Аннотация. Мақалада нақты уақыт режимінде ұшқышсыз ұшу

аппараттары (ҰҰА) мен құстарды тану және жіктеу моделін YOLOv10 нейрондық желісі негізінде әзірлеу қарастырылады. Зерттеу тақырыбы ҰҰА-ның әртүрлі салаларда кеңінен қолданылуына және соған байланысты қауіпсіздік мәселелерінің өзектілігіне байланысты маңызды болып отыр. Модельді оқыту үшін 7580 бейнеден тұратын деректер жиыны құрылды. Бұл бейнелер авторлық ҰҰА ұшу жазбаларынан және Roboflow, Ultralytics HUB, Kaggle сияқты ашық көздерден жиналды. Деректерді аннотациялау, аугментациялау және оларды оқыту, тексеру және тестілеу жиынтықтарына бөлу Roboflow платформасы арқылы жүзеге асырылды. Модель NVIDIA GeForce RTX 4080 графикалық процессоры негізінде Ultralytics құрылымын пайдаланып оқытылды. Нәтижесінде алынған модель mAP50 және mAP50-95 көрсеткіштері бойынша YOLO-ның алдыңғы нұсқаларынан асып түсетін жоғары дәлдікті көрсетті. Сонымен қатар, бұл модель объектілерді сегменттеу және бақылау тапсырмаларын жоғары тиімділікпен орындай алатынын көрсетті, бұл оны оптикалық-электрондық бақылау жүйелері үшін перспективалы шешім етеді. Зерттеу нәтижелері ҰҰА мен құстарды анықтау жүйелерін әзірлеушілер үшін пайдалы болып, стратегиялық маңызды нысандардағы қауіпсіздікті арттыруға ықпал етуі мүмкін.

Түйін сөздер: нейрондық желілер, жіктеу, тану, оптикалық электронды бақылау арналары, UAV, YOLO, конволюциялық нейрондық желілер

Дәйексөздер үшін: В.В. Семенюк, И.Г. Курмашев, В.В. Сербин, Л.Б. Курмашева, А.Е.Адильбеков. YOLOV10 нейрлік желілігінің негізінде нақты уақытта ұшқышсыз авиацияны тану моделін әзірлеу // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 320–332 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.019.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА МОДЕЛИ РАСПОЗНАВАНИЯ БПЛА В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ НА ОСНОВЕ НЕЙРОННОЙ СЕТИ YOLOV10

В.В. Семенюк[1], И.Г. Курмашев[1], В.В. Сербин[2], Л.Б. Курмашева[1], А.Е. Адильбеков[1]

[1]НАО «Северо-Казахстанский университет имени М. Козыбаева», Петропавловск, Казахстан;

[2]Казахский национальный исследовательский технический университет им. К.И. Сатпаева, Алматы, Казахстан.

E-mail: vvsemenyuk@ku.edu.kz

Семенюк В.В. — магистр технических наук, НАО «Северо-Казахстанский университет имени М. Козыбаева», Петропавловск, Казахстан
E-mail: vvsemenyuk@ku.edu.kz, https://orcid.org/0000-0002-8580-7326;

Курмашев И.Г. — кандидат тех. наук, НАО «Северо-Казахстанский университет имени М. Козыбаева», Петропавловск, Казахстан
E-mail: ikurmashev@ku.edu.kz, https://orcid.org/0000-0001-9872-7483;
Сербин В.В. — кандидат технических наук, ассоциированный профессор, Казахский национальный исследовательский технический университет им. К.И. Сатпаева, Алматы, Казахстан
E-mail: v.serbin@satbayev.university; https://orcid.org/0000-0002-5807-3873;
Курмашева Л.Б. — магистр технических наук, НАО «Северо-Казахстанский университет имени М. Козыбаева», Петропавловск, Казахстан
E-mail: lbkurmasheva@ku.edu.kz, https://orcid.org/0000-0001-5392-5873;
Адильбеков А.Е. — магистр технических наук, НАО «Северо-Казахстанский университет имени М. Козыбаева», Петропавловск, Казахстан
E-mail: Aeadilbekov@ku.edu.kz, https://orcid.org/0000-0002-2658-8792.

Аннотация. В статье представлена разработка модели распознавания и классификации беспилотных летательных аппаратов (БПЛА) и птиц в режиме реального времени на основе нейронной сети YOLOv10. Актуальность исследования обусловлена растущим использованием БПЛА в различных сферах и возникающими, в связи с этим вопросами обеспечения безопасности. Для обучения модели был сформирован датасет из 7580 изображений, собранных как из собственных архивов полётов БПЛА, так и из открытых источников, включая Roboflow, Ultralytics HUB и Kaggle. Аннотирование, аугментация и разбиение данных на обучающую, валидационную и тестовую выборки были выполнены с использованием платформы Roboflow. Обучение модели проводилось на графическом процессоре NVIDIA GeForce RTX 4080 с применением фреймворка Ultralytics. Полученная модель показала высокую точность, превзойдя предыдущие версии YOLO по метрикам mAP50 и mAP50-95. Также была продемонстрирована высокая эффективность в задачах сегментации и трекинга объектов, что делает её перспективным решением для систем оптикоэлектронного наблюдения. Результаты исследования могут быть полезны разработчикам систем обнаружения БПЛА и птиц и способствовать повышению уровня безопасности на охраняемых объектах.

Ключевые слова: нейронные сети, классификация, распознавание, оптикоэлектронные каналы наблюдения, БПЛА, YOLO, сверточные нейронные сети

Для цитирования: В.В. Семенюк, И.Г. Курмашев, В.В. Сербин, Л.Б. Курмашева, А.Е. Адильбеков. Разработка модели распознавания бпла в режиме реального времени на основе нейронной сети YOLOV10 // Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 320–332. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.019.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Real-time recognition of unmanned aerial vehicles (UAVs) through surveillance cameras using neural networks and visual detectors is an urgent area for scientific research. This is driven by the need for public and state security, as UAVs are now being actively used for illegal activities including espionage, smuggling delivery, and combat raids. UAV recognition detection systems are designed to detect such threats in a timely manner and respond appropriately to neutralize them. They also play a significant role in protecting critical infrastructure such as airports, power plants and government buildings where there is a substantial risk of potential attacks.

Real-time drone recognition helps enforce flight regulations and control restricted areas. This is especially important when there are strict requirements and restrictions on the use of UAVs in certain areas, such as military bases and private territories. Modern neural networks and machine learning algorithms can process copious amounts of data, which provides high accuracy recognition and opens new opportunities for developing effective monitoring systems. Such systems can be integrated with existing video surveillance systems, making them more affordable and efficient. The algorithms used as software modules for UAV detection and recognition in optoelectronic surveillance channels are YOLO, Faster R-CNN, and SSD. In (Alkentar et all., 2021: 19–31), the authors found that single-stage YOLO detectors and two-stage Faster R-CNN algorithms compared to SSD provide better accuracy of UAV recognition, while the latter algorithm is also effective in target detection tasks. However, according to the research results presented in (Seidaliyeva et all., 2020; Singha et all., 2021). trained YOLO models outperform Faster RCNN in terms of accuracy metrics and speed (frames per second - FPS), which makes the former algorithm more promising for real-time UAV and bird recognition and classification tasks. Considering the recent advances in the improvement of UAV recognition models based on the YOLO algorithm, we should mention the works (Seidaliyeva et all., 2020; Muzammul, et all., 2024: 1–1). In (Seidaliyeva et all., 2020), the authors prepared a dataset in the form of UAV images for training the YOLOv2 neural network. As a result of training, the model reached the maximum value of mAP50 (average accuracy at the threshold of intersection of bounding boxes 50 %) - 0.75. In (Alsanad et all., 2022), the authors studied the features of UAV recognition by YOLOv3 neural network models. By evaluating the qualitative performance of the trained experimental model, it is proposed to introduce densely connected modules to improve the inter-layer connectivity of convolutional neural networks, thereby improving the accuracy and FPS. As a result, the proposed model demonstrated mAP50-95 values of 0.36 and 60 FPS, which outperform the original YOLOv3 model by 0.03 and 24.45 FPS, respectively. The mAP values presented in (Seidaliyeva et all., 2020; Alsanad et all., 2022) are not equivalent, as mAP50-95 (Alsanad et all., 2022) defines the average accuracy at the

95 % bounding box crossing threshold, which implies comparatively lower values compared to mAP50. A better model of the YOLO architecture, YOLOv4, is used in (Singha et all., 2021), where the authors achieved mAP50 values of 0.75, which is consistent with the results of (Seidaliyeva et all., 2020). However, the YOLOv4 models can be considered preferable to the YOLOv2 algorithm primarily due to its higher FPS, which is important for recognizing small objects such as UAVs in real time. In (Al-Qubaydhi et all., 2022; Aydin et all., 2023), YOLOv5 was used as a pre-trained neural network, which is currently considered to be the most popular model of the YOLO architecture for solving the problems of recognizing and classifying any objects from a user's dataset, with training and inference available on the CPU. The neural network from the first paper (Al-Qubaydhi et all., 2022; Aydin et all., 2023), achieved a mAP50 value of 0.947, while the authors from trained the model to a mAP50 value of 0.904. The paper (Zhai et all., 2023) presents the results of training more advanced versions of YOLO:

- YOLOv5: mAP50 - 0.912;
- YOLOX: mAP50 - 0.887;
- YOLOv7: mAP50 - 0.524;
- YOLOv7-tiny: mAP50 - 0.85;
- YOLOv8 - 0.953.

However, the authors [7] did not provide the performance of the mAP50-95 metric, which can provide more reliable information about the quality of UAV recognition and detection. In 2023, Ultralytics presented an improved algorithm, YOLOv9 (Li et all., 2024). By incorporating programmable gradient information (PGI) and an efficient layer aggregation network into the architecture, this model will limit the loss of information in successive layers of a deep neural network. YOLOv9 like previous versions utilizes a non-maximal suppression (NMS) algorithm in the post-processing stage, designed to remove unnecessary object bounding boxes. Its use often discards useful predictions and increases computational cost, which is detrimental to the detection and recognition performance of small and maneuverable objects such as UAVs and birds. In this paper, a YOLOv10 model will be trained, which by incorporating new post-processing techniques from the DETR architecture (Muzammul, et all., 2024: 1–1), is hypothesized to increase the recognition and classification accuracy of UAVs and birds. As an analytical accuracy parameter that guarantees the effectiveness of the developed model, mAP50-95 is chosen.

Materials and methods

A more advanced model for real-time recognition and classification of UAVs and birds is based on the YOLOv10 algorithm. YOLOv10 is a neural network based on two key principles: training the model without NMPs and designing based on efficiency and accuracy. The first concept is provided by integrating the advantages of neural networks of the DETR architecture, namely the use of dual label assignment and a consistent metric for matching predictions (Figure 1).

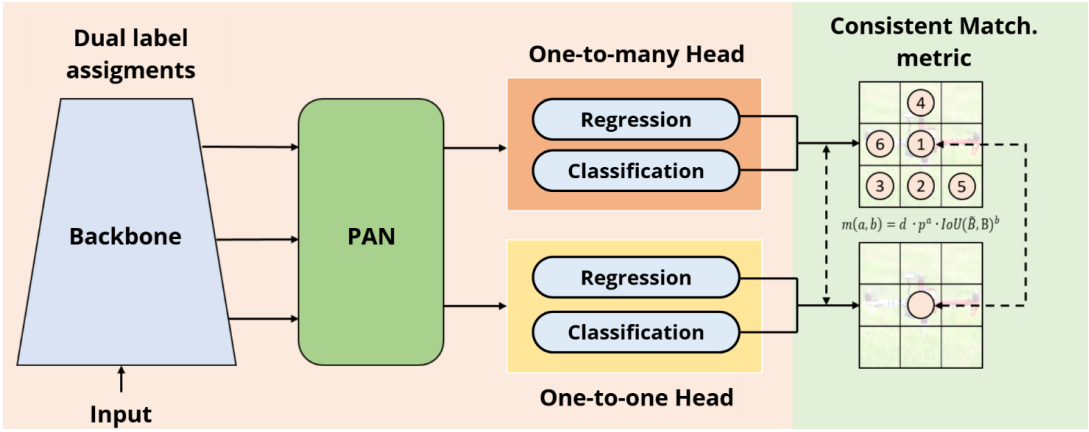


Fig. 1. Concept of the YOLOv10 Model

The dual purpose of labels is defined by the architectural solution of using two Heads, where the first Head performs One-to-One matching (provides only one prediction to each benchmark from the user dataset without the use of NMPs) and the second Head performs One-to-Many matching. By utilizing the One-to-Many branch, comprehensive and tight control is provided to ensure that accuracy is maintained. Throughout the training phase, both Heads are used simultaneously, except for the Inference (prediction) phase. At this stage, only the One-to-One branch is used, this decision is made to reduce computational cost while maintaining the accuracy and efficiency of the model. To guarantee the coincidence of the forecasts of the two branches, a consistent metric (or matching metric) is used. The principle of operation of this metric is described by equation (1):

$$m(a,b) = d \cdot p^a \cdot IoU(\tilde{B},B)^b \quad (1)$$

where m is the corresponding accuracy metric ("One to One" or "One to Many"), d is the degree of localization of the prediction reference point within the instance, p is the classification score, a and b are the parameters defining the classification and localization tasks, respectively, IoU is the parameter defining the area of overlap between the predicted and actual frames, is the bounding boxes of the prediction and the instance, respectively. Matching the metrics of both branches ensures that the best samples are matched, thus comprehensive learning control is realized. Designing a performance-based model involves facilitating the classification channel by using two depth-separable convolutions followed by a 1x1 convolution (point convolution). The point convolution allows increasing the number of channels, while the depth convolution reduces the spatial dimensions, such a solution allows reducing the computational cost, as well as preserving more useful information in the downsampling stage. The accuracy orientation of the model is ensured by increasing the dimensionality of kernels in deep convolution layers, as well as by using the partial self-awareness module (PSM).

YOLOv10 supports 6 scales depending on the limitations of the computational resources of the hardware. The YOLOv10m version, which represents a medium version for universal use, is chosen for the experimental study. The training dataset used for the YOLOv10m model consists of 7580 images, comprising two balanced classes: UAVs and birds. The dataset was assembled from a combination of open-source repositories (Roboflow, Ultralytics HUB, Kaggle) and proprietary UAV flight footage collected in field conditions. The selection process deliberately incorporated diverse environmental conditions, including different lighting levels, backgrounds, and object scales, to ensure the model's robustness and generalizability. To further improve variability and mitigate overfitting, a custom Python-based augmentation script was developed. It performed operations such as horizontal and vertical flipping, rotation, blurring, contrast and brightness adjustments, noise addition, and scaling transformations. These augmentations were applied selectively based on class and environmental context. The dataset was then annotated, augmented, and split into training, validation, and test sets in a 70/20/10 ratio using the Roboflow platform.

Training, validation and testing of the model (Fig.2) YOLOv10m is realized based on AD103 graphics processor of NVIDIA GeForce RTX 4080 graphics card with support of CUDA Toolkit 12.1. The program code in Python language is implemented in PyCharm 2024 environment using Ultralytics framework, which allows using the open-source resource of pre-trained YOLOv10 models. The following hyperparameters are set for training the neural network on the user set: number of epochs: 300; packet size: 16; learning rate: 0.001; momentum: 0.9; weight drop: 0.0005 and image size: 640. The trained YOLOv10m model has a file size of "best.pt" of 101 MB. To determine the FPS, the trained neural network is tested on inference using two test videos of UAV and bird flights.

To compare the accuracy of the experimental neural network with the previous versions, the YOLOv8m and YOLOv9m models corresponding in purpose were also trained on the user dataset. The mAP50-95 metric was chosen as the comparative accuracy parameter to evaluate the ability of visual detectors to recognize and localize small objects.

Results

Figure 3a-d shows the accuracy metrics of the YOLOv10m model trained on a custom dataset of UAVs and birds.

The frames of testing the trained model on the inferno using two videos of a DJI F450 UAV flight and a bird are shown in Figure 4a,b.

Figure 5a, b shows the comparison plots of mAP50-95 accuracy metric and fast performance (FPS) of YOLOv8m, YOLOv9m and YOLOv10m neural networks.

As a result of the diagrams analysis, we can conclude about the effectiveness of using the trained YOLOv10m model in the tasks of recognition and classification of UAVs and birds in optoelectronic surveillance systems.

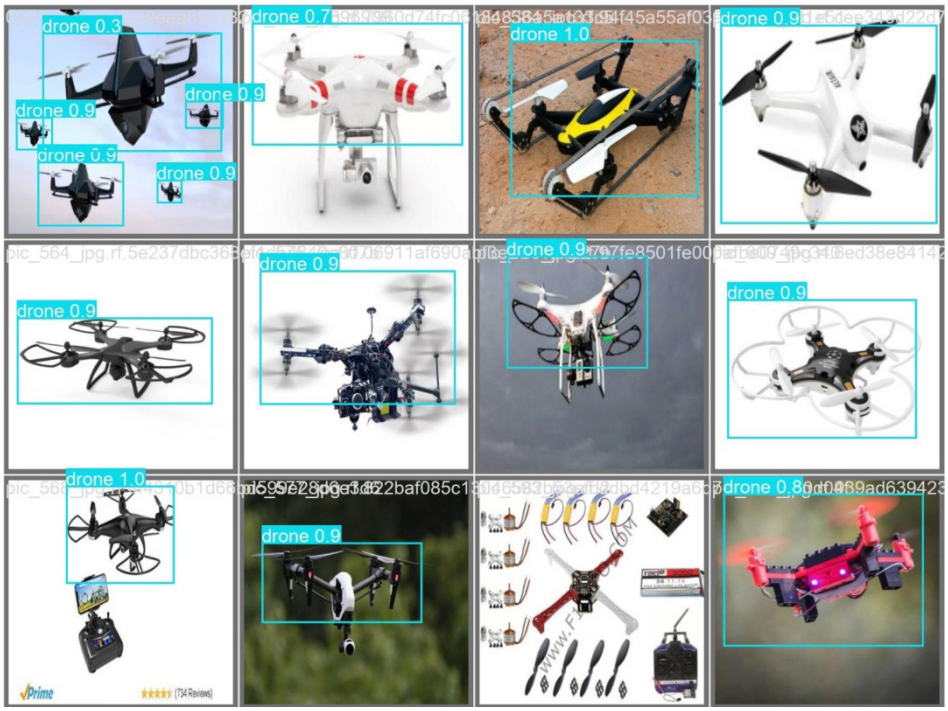


Fig. 2. Testing of the Trained YOLOv10m Neural Network Model

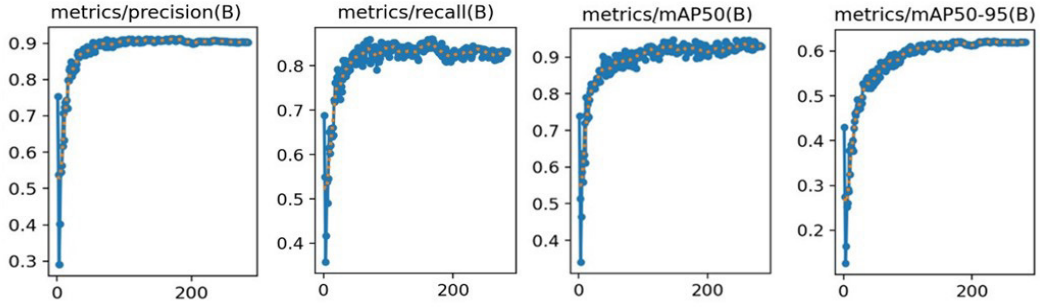


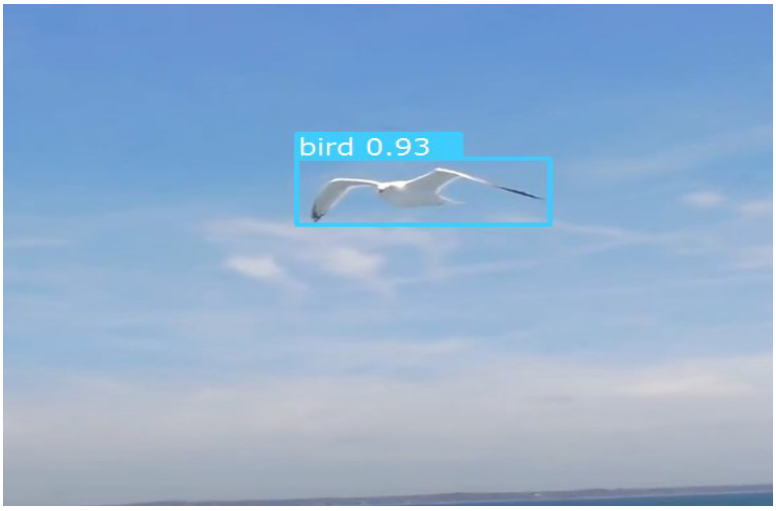
Fig. 3. Metrics of Training Results of the YOLOv10m Neural Network: a - Precision; b - Recall; c - mAP50; d - mAP50-95.

Discussion

A dataset of 7580 UAV and bird images is prepared for training the experimental model of YOLOv10m neural network. After annotation and augmentation stages, the user dataset using Roboflow.com service is distributed in the percentage of 70/20/10 for training, validation and testing tasks, respectively. The training of experimental neural network resulted in the following maximum values of accuracy metrics (Fig. 4a-d):

- Completeness: 0.883;
- Accuracy: 0.907;
- mAP50: 0.953;
- mAP50-95: by 0.628.

a



b

Fig. 4. Inference Frames of the Trained YOLOv10m Model: (a) Frame from the DJI F450 Flight Video; (b) Frame of a Bird (Seagull) Flight

In terms of mAP50 metric, this model outperforms detectors from (Seidaliyeva et all., 2020; Alsanad et all., 2022; Singha et all., 2021; Al-Qubaydhi et all., 2022; Aydin et all., 2023; Zhai et all., 2023). Comparing the mAP50-95 metrics (Fig. 5a), YOLOv10m outperforms YOLOv8 by 0.058 and YOLOv9 by 0.031 in terms of accuracy, which proves the greater efficiency of the first model to recognize and localize small objects such as UAVs and birds in the far-field surveillance area. In terms of FPS, YOLOv10m is second only to the trained model YOLOv8m, which proves that the first model can be used in real-time in optoelectronic surveillance channels, combining the segmentation and tracking functions of small objects such as UAVs and birds.

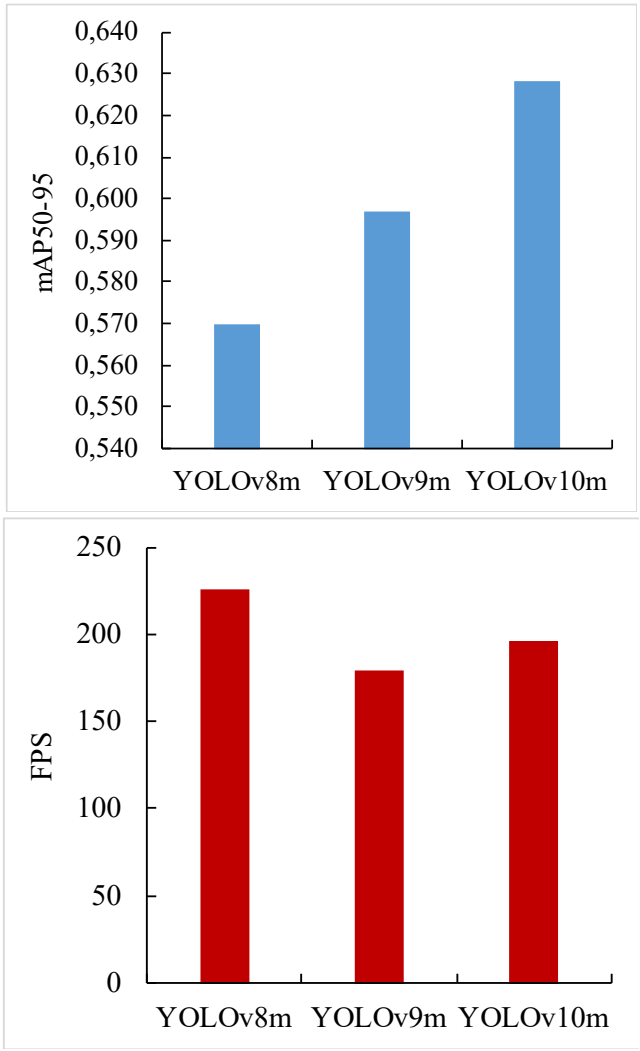


Fig. 5. Comparison Plots of mAP50-95 Accuracy Metrics and FPS of YOLOv8m, YOLOv9m and YOLOv10m Neural Networks

The results of the experiments indicate that the trained YOLOv10m model demonstrates strong robustness to varying environmental and operational conditions. The dataset included images captured under different weather scenarios (cloudless sky, partial cloudiness, low light, and fog-like conditions), various camera angles, and dynamic backgrounds. As a result, the model maintained high detection performance across these variations, confirming its adaptability to real-world deployments in outdoor optoelectronic surveillance systems. Furthermore, this opens avenues for future research focused on expanding the model's capabilities in more complex environments, such as night-time recognition, detection under severe weather disturbances, and integration with radar or infrared sensor data for multi-modal UAV classification.

Conclusion

1) In order to develop a new efficient model for UAV and bird recognition and classification, the YOLOv10m model is selected, which bypasses the Non-Maximal Suppression (NMS) algorithm in its architecture by introducing the concepts of dual label assignment, consistent metric for prediction matching and focus on accuracy and efficiency in design. A user dataset of 7580 images is prepared to train the experimental models, of which 5306 (70 %) are specifically designed for training, 1516 (20 %) are oriented for validation and the remaining 758 (10 %) are used in the testing phase of the neural network.

2) YOLOv10m training is implemented based on AD103 GPU of NVIDIA GeForce RTX 4080 video card with CUDA Toolkit 12.1 support. As a result of evaluation of the obtained accuracy metrics, YOLOv10 outperforms the known visual detectors YOLOv2, YOLOv3, YOLOv4, YOLOv5, YOLOv7, YOLOX and YOLOv8 described in (Seidaliyeva et all., 2020; Alsanad et all., 2022; Singha et all., 2021; Al-Qubaydhi et all., 2022; Aydin et all., 2023; Zhai et all., 2023) by mAP50.

3) To compare the efficiency of recognition and localization of small objects using the mAP50-95 metric, previous versions of the YOLO algorithm: YOLOv8 and YOLOv9 were trained on a user dataset. YOLOv10m was 9% more accurate than the former and 5% more accurate than the latter.

4) As a result of testing the trained YOLOv10m neural network on inference, it is found that this model can recognize and classify UAVs and birds in real time with high accuracy and efficiency. According to the high FPS values (average 196 FPS), this model is capable of segmenting and tracking localized objects, combining the classification task.

5) The results of this study will be useful to the developers of UAV and bird detection, recognition and classification systems.

REFERENCES

Alkentar, Saad & Alsahwa, B. & Assalem, A. & Karakolla, D. (2021). Practical comparation of the accuracy and speed of YOLO, SSD and Faster RCNN for drone detection. *Journal of Engineering.* — 27. — 19-31. 10.31026/j.eng.2021.08.02.

Alsanad, Hamid & Sadik, Amin & Ucan, Osman & Ilyas, Muhammad & Bayat, Oguz. (2022). YOLO-V3 based real-time drone detection algorithm. Multimedia Tools and Applications. — 81. — 1–14. 10.1007/s11042-022-12939-4.

Al-Qubaydhi, Nader & Alenezi, Abdulrahman & Alanazi, Turki & Senyor, Abdulrahman & Alanezi, Naif & Alotaibi, Bandar & Alotaibi, Munif & Razaque, Abdul & Abdelhamid, Abdelaziz & Alotaibi, Aziz. (2022). Detection of Unauthorized Unmanned Aerial Vehicles Using YOLOv5 and Transfer Learning. Electronics. — 11. — 2669. 10.3390/electronics11172669.

Aydin B., Singha S. (2023). Drone Detection Using YOLO v5. Eng. — 4. DOI: 10.3390/eng4010025.

Li, Jun & Feng, Yongqiang & Shao, Yanhua & Liu, Feng. (2024). IDP-YOLOV9: Improvement of Object Detection Model in Severe Weather Scenarios from Drone Perspective. Applied Sciences. — 14. — 5277. 10.3390/app1412125277.

Muzammul, Muhammad & Algarni, Abdul & Ghadi, Yazeed & Assam, Muhammad. (2024). Enhancing UAV Aerial Image Analysis: Integrating Advanced SAHI Techniques with Real-Time Detection Models on the VisDrone Dataset. IEEE Access. — Pp. 1–1. 10.1109/ACCESS.2024.3363413.

Seidaliyeva U., Alduraibi M., Ilipbayeva L., Almagambetov A. (2020). Detection of Loaded and Unloaded UAV Using Deep Neural Network. In Proceedings of the 2020 Fourth IEEE International Conference on

Ro-botic Computing (IRC), Taichung, Taiwan. — November 9–11. — 2020. — Pp. 490–494. DOI: 10.1109/IRC.2020.00093

Singha S., Aydin B. (2021). Automated Drone Detection Using YOLO v4. Drones. September —2021. DOI: 10.3390/drones5030095.

Zhai X., Huang Z., Li T., Liu H., Wang S. (2023). YOLO-Drone: An Optimized YOLOv8 Network for Tiny UAV Object Detection. Electronics. — 12. —3664. DOI: 10.3390/electronics12173664

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 333–350
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.020

МРНТИ 28.23.37

DETECTING DUPLICATES IN KAZAKH TEXTS: A COMPARISON OF TF-IDF, WORD AND SENTENCE EMBEDDINGS

A.O. Tleubayeva*, S.V. Biloshchytska, O. Kuchanskyi, A.A. Mukhatayev, A.B. Nugumanova
Astana IT University, Astana, Kazakhstan.
E-mail: bsv@astanait.edu.kz

Arailym Tleubayeva — PhD student, senior lecturer at the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
E-mail: a.tleubayeva@astanait.edu.kz, https://orcid.org/0000-0001-9560-9756;
Svitlana Biloshchytska — Doctor of Technical Sciences, Professor at the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-0856-5474;
Oleksandr Kuchanskyi — Doctor of Technical Sciences, Professor at the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0003-1277-8031;
Aidos Mukhatayev, Candidate of Pedagogical Sciences, Professor at the School of General Education Disciplines, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-8667-3200;
Aliya Nugumanova— PhD, Head of the Scientific and Innovation Center «Big Data and Blockchain Technologies», Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-5522-4421.

Abstract. This paper presents a comprehensive comparison of TF-IDF, word, and multilingual sentence embeddings for automatic duplicate detection in Kazakh texts. Experiments use the KazakhTextDuplicates dataset with labels for exact, paraphrase, contextual, and partial duplicates. All models were evaluated within a unified setup featuring standardized preprocessing, L2-normalized vectors, and validation-based threshold tuning. The Word2Vec model with TF-IDF weighting achieved the highest performance (F1 = 0.996; ROC-AUC = 0.9999; PR-AUC = 0.9999). The TF-IDF (1–3-grams) method remained competitive for exact and partial overlaps (PR-AUC = 0.932; ROC-AUC = 0.775), while FastText provided the best recall (R ≈ 0.99) at moderate precision. Among multilingual models, BGE-m3

and Snowflake Arctic achieved the best PR-AUC (≈0.614). In retrieval, the BM25 followed by dense re-ranking pipeline produced a small but consistent improvement over dense-only search (Recall@10: +0.04–0.12 pp; nDCG@10: +0.10–0.13 pp), confirming the effectiveness of combining lexical and semantic features for duplicate detection in morphologically rich, low-resource languages such as Kazakh.

Keywords: duplicate detection; Kazakh language; TF-IDF; word embeddings; sentence embeddings; semantic similarity; BM25; dense retrieval; hybrid reranking; low-resource NLP

For citation: A.O. Tleubayeva, S.V. Biloshchytsk***a***, O. Kuchanskyi, A.A. Mukhatayev, A.B. Nugumanov***a***. Detecting duplicates in kazakh texts: a comparison of tf-idf, word and sentence embeddings // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 333–350. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.020.

Conflict of interest: The authors declare that there is no conflict of interest.

Funding Information: This research work was carried out within the framework of the scientific project AP23490123 «Development of a system to detect plagiarism using combined methods and models for finding near-duplicate, focusing on the Kazakh language.» for 2024-2026, financed by the Committee of Science Ministry of Science and Higher Education of the Republic of Kazakhstan.

ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДЕРДЕГІ ДУБЛИКАТТАРДЫ АНЫҚТАУ: TF-IDF, СӨЗ ЖӘНЕ СӨЙЛЕМ ЭМБЕДДИНГТЕРІН САЛЫСТЫРУ

А.О. Тлеубаева*, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова

Astana IT University, Astana, Kazakhstan.

E-mail: a.tleubayeva@astanait.edu.kz

Тлеубаева Арайлым — PhD докторант, «Жасанды интеллект және деректер ғылымы» мектебінің сеньор-лекторы, Astana IT University, Астана, Қазақстан
E-mail: a.tleubayeva@astanait.edu.kz, https://orcid.org/0000-0001-9560-9756;

Белошицкая Светлана — техника ғылымдарының докторы, «Жасанды интеллект және деректер ғылымы» мектебінің профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-0856-5474;

Кучанский Александр — техника ғылымдарының докторы, «Жасанды интеллект және деректер ғылымы» мектебінің профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0003-1277-8031;

Мухатаев Айдос — Педагогика ғылымдарының кандидаты, «Жалпы білім беру пәндері» мектебінің профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-8667-3200;

Нугуманова Алия — PhD, «Big Data and Blockchain Technologies», ғылыми-ин-новациялық орталығының жетекшісі, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-5522-4421.

Аннотация. Мақалада қазақ тіліндегі мәтін дубликаттарын автоматты түрде анықтау үшін TF-IDF, сөздік және көптілді сөйлем эмбеддингтері кешенді түрде салыстырылды. Эксперименттер KazakhTextDuplicates деректер жинағында жүргізілді, мұнда жұптар «нақты», «парафраз», «контекстік» және «ішінара» дубликат ретінде таңбаланған. Барлық модельдер бірыңғай эксперименттік ортада бағаланды: стандартталған алдын ала өңдеу, L2-нормаланған векторлар және валидация арқылы шек мәнін баптау. TF-IDF-пен салмақталған Word2Vec моделі ең жоғары нәтижелерге жетті (F1 = 0.996; ROC-AUC = 0.9999; PR-AUC = 0.9999). TF-IDF (1–3-грамма) әдісі нақты және ішінара сәйкестіктер үшін тиімді болды (PR-AUC = 0.932; ROC-AUC = 0.775), ал FastText жоғары толықтық (R ≈ 0.99) көрсетті. Көптілді модельдер арасында BGE-m3 және Snowflake Arctic PR-AUC бойынша үздік нәтижелерге (≈0.614) жетті. Іздеу міндетінде BM25 және кейінгі тығыз қайта ранжирлеу тәсілі тығыз іздеумен салыстырғанда аз болса да тұрақты өсім көрсетті (Recall@10: +0.04–0.12 п.б.; nDCG@10: +0.10–0.13 п.б.), бұл лексикалық және семантикалық белгілерді біріктірудің тиімділігін дәлелдейді.

Түйін сөздер: дубликаттарды анықтау; қазақ тілі; TF-IDF; сөздік эмбеддингтер; сөйлемдік эмбеддингтер; семантикалық ұқсастық; BM25; тығыз іздеу (dense retrieval); гибридті қайта ранжирлеу; ресурсы шектеулі тілдер үшін NLP

Дәйексөздер үшін: А.О. Тлеубаева, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова. Қазақ тіліндегі мәтіндердегі дубликаттарды анықтау: tf-idf, сөз және сөйлем эмбеддингтерін салыстыру // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 333–350 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.020.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ОБНАРУЖЕНИЕ ДУБЛИКАТОВ В КАЗАХСКИХ ТЕКСТАХ: СРАВНЕНИЕ TF-IDF, ЭМБЕДДИНГОВ СЛОВ И ЭМБЕДДИНГОВ ПРЕДЛОЖЕНИЙ

А.О. Тлеубаева*, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова
Astana IT University, Astana, Kazakhstan.
E-mail: a.tleubayeva@astanait.edu.kz

Тлеубаева Арайлым — PhD докторант, сеньор-лектор «Школы искусственного интеллекта и науки о данных», Astana IT University, Астана, Казахстан
E-mail: a.tleubayeva@astanait.edu.kz, https://orcid.org/0000-0001-9560-9756;
Белощицкая Светлана — доктор технических наук, профессор «Школы искусственного интеллекта и науки о данных», Astana IT University, Астана, Казахстан**.**

https://orcid.org/0000-0002-0856-5474;
Кучанский Олександр — доктор технических наук, профессор «Школы искусственного интеллекта и науки о данных», Astana IT University, Астана, Казахстан.
https://orcid.org/0000-0003-1277-8031;
Мухатаев Айдос — кандидат педагогических наук, профессор «Школы общеобразовательных дисциплин», Astana IT University, Астана, Казахстан
https://orcid.org/0000-0002-8667-3200;
Нугуманова Алия — PhD, директор НИЦ Big Data and Blockchain Technologies, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-5522-4421.

Аннотация. В статье представлен всесторонний сравнительный анализ методов TF-IDF, словарных и многоязычных эмбеддингов предложений для автоматического обнаружения дубликатов в казахских текстах. Эксперименты проведены на датасете KazakhTextDuplicates, включающем пары с метками «точный», «парафраз», «контекстуальный» и «частичный» дубликат. Все модели оценивались в единой экспериментальной схеме с унифицированной предобработкой, L2-нормированными векторными представлениями и подбором порога по валидационной выборке. Модель Word2Vec с TF-IDF-взвешиванием показала наилучшие результаты (F1 = 0.996; ROC-AUC = 0.9999; PR-AUC = 0.9999). Метод TF-IDF (1–3-граммы) продемонстрировал высокую точность для точных и частичных совпадений (PR-AUC = 0.932; ROC-AUC = 0.775), тогда как FastText обеспечил максимальную полноту (R ≈ 0.99) при умеренной точности. Среди многоязычных моделей лучшие показатели PR-AUC (≈0.614) получены для BGE-m3 и Snowflake Arctic. В задаче поиска дубликатов гибридная схема BM25 с последующим плотным переранжированием обеспечила небольшой, но стабильный прирост по сравнению с плотным поиском (Recall@10: +0.04–0.12 п.п.; nDCG@10: +0.10–0.13 п.п.), что подтверждает эффективность сочетания лексических и семантических признаков для морфологически сложных, низкоресурсных языков.

Ключевые слова: обнаружение дубликатов; казахский язык; TF-IDF; эмбеддинги слов; эмбеддинги предложений; семантическое сходство; BM25; плотный поиск (dense retrieval); гибридный переранжиринг; NLP для низкоресурсных языков

Для цитирования: А.О. Тлеубаева, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова. Обнаружение дубликатов в казахских текстах: сравнение tf-idf, эмбеддингов слов и эмбеддингов предложений//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 333–350. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.020.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

The rapid expansion of digital text corpora has intensified the demand for effective tools in information retrieval, plagiarism detection, and large-scale document management. A fundamental challenge in this area is duplicate detection, which refers to identifying exact or near-exact repetitions of text fragments across collections. While this task has been extensively studied for high-resource languages such as English, Russian, and Chinese (Wang et al., 2020; Li et al., 2021), there remains a considerable gap for low-resource languages, including Kazakh. Kazakh poses unique challenges for duplicate detection due to its agglutinative morphology, rich inflection, and free word order (Mussiraliyeva et al., 2024). Small variations in affixes or syntax often alter surface forms without affecting semantic meaning, complicating the design of robust detection systems.

Traditional statistical approaches, particularly term frequency–inverse document frequency (TF-IDF), have proven effective for identifying near-exact duplicates but struggle with paraphrased or contextually similar fragments (Cheng et al., 2020). Advances in neural embeddings have significantly expanded the scope of text similarity modeling. Word embeddings such as FastText can capture subword information and morphological patterns, which is especially beneficial for agglutinative languages (Bojanowski et al., 2017). Building upon this, sentence embeddings trained on large multilingual corpora (e.g., LaBSE (Feng et al., 2020), E5 (Wang et al., 2024), BGE (Chen et al., 2024), GTE/mGTE (Zhang et al., 2024), Snowflake Arctic (Yu et al., 2024), and Alibaba GTE) enable robust comparison of entire sentences or paragraphs, encoding both syntactic and semantic similarity. These models have recently achieved strong performance in multilingual semantic similarity and retrieval tasks (Xu et al., 2025; Mansurova et al., 2024).

Despite these advances, Kazakh remains underrepresented in large-scale pretraining datasets, and the applicability of embedding-based approaches to duplicate detection in Kazakh has not been systematically investigated. Prior work in Kazakh NLP has focused mainly on morphological analysis, corpus construction, and part-of-speech tagging (Akhmed-Zaki et al., 2021; Mansurova & Rakhimova, 2025), with only limited exploration of semantic similarity or duplicate detection. To the best of our knowledge, no comparative study has benchmarked statistical, word-level, and sentence-level representations on a dedicated Kazakh duplicate detection dataset.

This study addresses this gap by conducting a comparative evaluation of TF-IDF, word embeddings, and sentence embeddings for duplicate detection in Kazakh texts. We employ the KazakhTextDuplicates dataset (Tleubayeva, 2025), which includes labeled examples across categories such as exact duplicates, paraphrases, contextual duplicates, and partial overlaps. Our research is guided by the following questions:

1.Which representation methods provide the highest accuracy for duplicate detection in Kazakh texts?

2. How robust are these methods to preprocessing and parameter variations?

3.Does a hybrid re-rank strategy improve retrieval-oriented recall compared to standalone methods?

Based on prior findings, we formulate the following hypotheses:

H1: Sentence embeddings will significantly outperform TF-IDF and word embeddings on paraphrased and contextual duplicates, as they capture semantic equivalence beyond surface similarity (Feng et al., 2020; Wang et al., 2024; Chen et al., 2024; Yu et al., 2024; Zhang et al., 2024).

H2: Character n-gram TF-IDF will remain competitive on exact and partial duplicates, where surface overlap dominates, even if its performance degrades on semantic cases (Cheng et al., 2020; Bojanowski et al., 2017).

H3: A hybrid BM25 followed by dense re-ranking strategy will achieve higher Recall@k than standalone BM25 or dense models, providing a balanced approach to both lexical and semantic similarity (Xu et al., 2025; Mansurova et al., 2024).

By systematically evaluating these approaches, this paper contributes to the development of duplicate detection systems for Kazakh and provides insights applicable to other morphologically rich, low-resource languages.

Materials and methods

Dataset and Preprocessing

We employ the publicly available KazakhTextDuplicates dataset (Tleubayeva, 2025), created specifically for duplicate detection in Kazakh. The corpus consists of pairs of text fragments annotated as duplicate or non-duplicate, with fine-grained duplicate categories that include exact, paraphrase, contextual, and partial overlaps. This label design enables separate analysis of surface-level and semantic similarity cases.

For fair comparison, we split the data into train/validation/test = 70%/10%/20% using stratified sampling by duplicate type, preserving the proportion of each category across subsets. Stratification is especially important to maintain balanced coverage of both trivial exact matches and more challenging paraphrased/contextual examples.

For the word-embedding baselines reported in this paper, we use an operational subset stored at kk_pairs_with_nondup.csv, derived from the original dataset. Each record contains the source text A (content), the modified text B (modified_content), and a label type_duplicate ∈ {exact, paraphrase, partial, nondup}. In this subset, contextual duplicates from the source corpus are not included, to keep a consistent set of positive types for per-type F1 versus the non-duplicate class used as the negative reference.

Given the agglutinative morphology of Kazakh, preprocessing was crucial for improving model robustness. The following steps were applied consistently across all methods:

1. Text normalization: lowercasing and Unicode normalization.

2. Stop-word removal: using a curated list of Kazakh stop-words.
3. Lemmatization/Stemming: leveraging available Kazakh morphological analyzers to reduce words to canonical forms.
4. Segmentation: texts were segmented into sentences or short paragraphs to form candidate fragments for comparison.

This uniform preprocessing ensured comparability between statistical, word-level, and sentence-level approaches.

Representation Methods

This study employed three main approaches to text representation: the statistical TF-IDF method, distributed word embeddings, and sentence embeddings (Table 1). Each of these techniques was implemented and tested within the task of duplicate detection in a corpus of Kazakh texts.

The TF-IDF (term frequency–inverse document frequency) method was used to construct vector representations of text fragments based on term frequency distributions. Several configurations were tested, including unigram, bigram, and trigram models, as well as character-level n-grams (char_wb) ranging from three to six characters. The latter was particularly relevant for Kazakh, a morphologically rich language with complex affixation. To improve representation quality, frequency thresholds were applied: extremely rare and overly frequent terms were excluded using min_df and max_df parameters. This setup enabled a direct comparison between lexical and character-based features, especially in detecting exact and partial duplicates (Li et al., 2022; Bakiyev, 2022).

Distributed word representations were built using pre-trained FastText and Word2Vec models adapted for multilingual and Kazakh corpora. Each word was mapped into a vector space, and fragment-level vectors were obtained by aggregating individual word vectors. Several aggregation strategies were considered: simple averaging, TF-IDF weighted averaging to highlight informative terms, and max pooling as an alternative baseline. These methods provided an intermediate level of representation between surface-level statistical features and context-sensitive sentence-level embeddings (Biloshchytska et al., 2025; Ayazbayev et al., 2023).

To capture semantic and contextual information at the sentence or paragraph level, several modern multilingual encoders from Hugging Face were evaluated. Specifically, the models included LaBSE (Feng et al., 2020), intfloat/multilingual-e5-base (Wang et al., 2024), BAAI/bge-m3 (Chen et al., 2024), Snowflake Arctic (Yu et al., 2024), and Alibaba GTE/mGTE (Zhang et al., 2024). Each fragment was mapped to a single vector, which was subsequently L2-normalized to ensure consistency in similarity computation. Where appropriate, dimensionality reduction was applied to standardize the representations. These models offered a more robust means of identifying paraphrased and contextual duplicates compared to purely statistical approaches.

Table 1. Summary of Methods and Hypotheses Tested.

Method Group	Representative Models / Settings	Target Duplicate Types	Related Hypothesis	Expected Role in Results
TF-IDF	Word n-grams (1–3), min_df = 3–5, max_df = 0.85–0.95; char_wb n-grams (3–6)	Exact duplicates, Partial overlaps	H2	Strong baseline for surface similarity; char_wb expected to remain competitive on near-exact cases
Word Embeddings	FastText, Word2Vec; pooling strategies: mean, TF-IDF-weighted, max	Partial overlaps, some paraphrases	(bridging case between H1 and H2)	Better than TF-IDF for morphologically rich fragments; less robust than sentence embeddings
Sentence Embeddings	LaBSE, multilingual-E5, BGE-m3, Snowflake Arctic, GTE	Paraphrases, Contextual duplicates	H1	Expected to outperform TF-IDF/WordEmb on semantic similarity tasks
Hybrid Retrieval	BM25 → Dense rerank (cosine similarity on embeddings)	All categories (retrieval scenario)	H3	Expected to improve Recall@k compared to standalone BM25 or dense retrieval

Evaluation Metrics

All sentence vectors are L2-normalized prior to similarity computation. The primary similarity measure is cosine similarity; for robustness, we additionally evaluate using Euclidean and Manhattan distances.

When using distances, either (i) convert them to a "pseudo-similarity" s=1−d, or (ii) tune the threshold directly in the distance scale over a meaningful range. On the validation split, the decision threshold τ is selected over a uniform grid (for cosine, $\tau \in [0.00, 0.99]$ with step 0.01). The optimal threshold τ^* is determined by maximizing the F1-score:

$$\tau^* = \arg\max_{\tau} F_1(\tau) \tag{1}$$

where $F1(\tau)$ denotes the F1-score computed at a given threshold τ.

To rigorously evaluate the performance of the duplicate detection methods, we employed both classification-oriented metrics and ranking-oriented metrics. These measures were chosen to capture complementary aspects of model performance: precision of duplicate detection, ability to retrieve all duplicates, robustness across thresholds, and performance on imbalanced data.

Precision quantifies the proportion of text pairs predicted as duplicates that are actually correct. Formally:

$$P = \frac{TP}{TP+FP} \tag{2}$$

where TP is the number of true positives (correctly identified duplicates), and FP is the number of false positives (non-duplicates misclassified as duplicates).

We use precision because in real-world scenarios (e.g., plagiarism detection or document management), false alarms can undermine trust in the system.

Recall measures the proportion of actual duplicates that were successfully retrieved by the model:

$$R = \frac{TP}{TP+FN} \tag{3}$$

where FN denotes false negatives (missed duplicates).

Recall is critical in our experiment because missing true duplicates reduces the effectiveness of applications such as information retrieval and knowledge deduplication in Kazakh corpora.

The F1-score balances precision and recall by computing their harmonic mean:

$$F1 = 2 \cdot \frac{P \cdot R}{P+R} \tag{4}$$

This metric is particularly suitable for our task, as it penalizes extreme trade-offs (e.g., high recall but very low precision). It provides a single score to compare methods under varying thresholds.

The area under the Receiver Operating Characteristic curve (ROC-AUC) evaluates the model's ability to discriminate between duplicates and non-duplicates across different thresholds:

$$ROC - AUC = \int_0^1 TPR(FPR)\, dFPR \tag{5}$$

where

$$TPR = \frac{TP}{TP+FN}, \quad FPR = \frac{FP}{FP+TN} \tag{6}$$

ROC-AUC is included for completeness as a widely recognized discrimination measure, though it may be less informative under strong class imbalance.

The area under the Precision–Recall curve (PR-AUC) is defined as:

$$PR - AUC = \int_0^1 P(R)\, dR \tag{7}$$

Unlike ROC-AUC, PR-AUC is more sensitive to imbalanced datasets. Since duplicate pairs are much rarer than non-duplicates in Kazakh corpora, PR-AUC provides a more realistic estimate of performance in practical scenarios.

Cosine similarity was used as the primary distance metric, as it is scale-invariant and robust to differences in vector magnitude:

$$CosineSim(x,y) = \frac{x \cdot y}{|x||y|} \tag{8}$$

For robustness testing, Euclidean and Manhattan distances were also included:

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \; d_{Manhattan}(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

(9)

Testing multiple distance functions ensured that results did not depend solely on a single similarity measure.

Results and discussion

In this experiment, models based on TF-IDF representations were evaluated using cosine similarity and L2 normalization of vectors. Two types of feature representations were considered: word n-grams and symbolic n-grams within the boundaries of words. The range of n-grams (from unigrams to trigrams), as well as minimum and maximum frequency thresholds, were varied for word models, which allowed six configurations to be formed. Symbolic models were used as control models and included four configurations with n-gram ranges from three to six characters.

The classification threshold was selected on a validation dataset using the grid search method in order to maximize the F1 metric. In all configurations, the optimal value was equal to $\tau^* = 0$. During testing, this resulted in completeness (Recall) = 1.0 and accuracy (Precision) ≈ 0.75, which roughly corresponds to the a priori proportion of the positive class in the sample. Consequently, the final value of F1 (≈0.859) reflected the class imbalance to a greater extent than the actual ability of the model to distinguish duplicates. This highlights the limitations of threshold metrics and the need to use quality rating measures such as ROC-AUC and PR-AUC.

As shown in Table 2, the best results were achieved when using word n-grams in the range (1, 3): PR-AUC ≈ 0.932 and ROC-AUC ≈ 0.775. The configuration with bigrams (1, 2) showed comparable, but slightly lower results (PR-AUC ≈ 0.930, ROC-AUC ≈ 0.767). At the same time, the model based only on unigrams showed a sharp decrease in quality (PR-AUC ≤ 0.625, ROC-AUC ≈ 0.205), which confirms the insufficiency of unigrams to reflect the context and morphological dependencies in the Kazakh language.

Control experiments with symbolic n-grams showed significantly worse results (PR-AUC ≈ 0.625, ROC-AUC ≈ 0.21), which indicates that such features, limited by word boundaries, are not able to effectively model interword morphological and paraphrased structures characteristic of the Kazakh language.

Table 2. Comparison of TF-IDF configurations during validation and test

Method (Analyzer)	N-gram	min_df	max_df	F1 (test)	Precision	Recall	ROC-AUC	PR-AUC	Comment
TF-IDF (word)	(1,2)	3–5	0.85–0.95	0.859	0.75	1.00	0.767	0.930	Stable result, optimal for near-exact duplicates
TF-IDF (word)	(1,3)	3–5	0.85–0.95	0.859	0.75	1.00	0.775	0.932	Best combination in terms of PR-AUC

TF-IDF (word)	(1,1)	3–5	0.85–0.95	0.859	0.75	1.00	0.205	0.624	Inferior performance in ROC/PR-AUC
TF-IDF (char_wb)	(3–6, 4–6)	3–5	0.85–0.95	0.859	0.75	1.00	0.210	0.625	Although F1 is equal, the curve-based metrics are weak

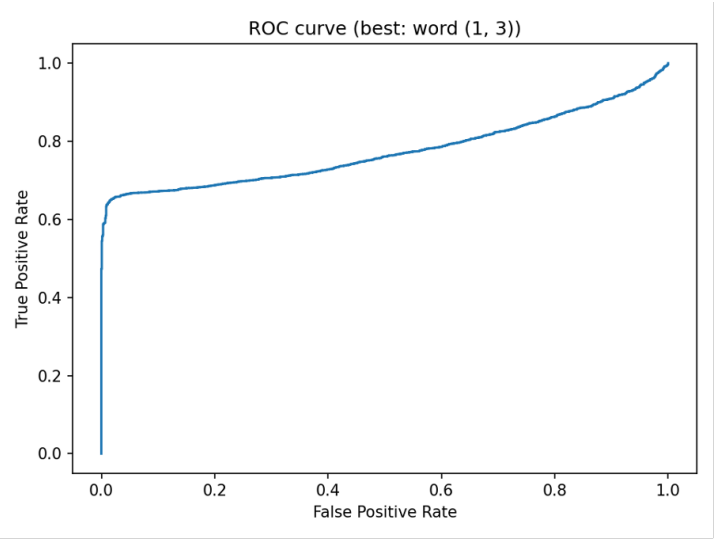


Fig.1. Receiver Operating Characteristic (ROC) curve for TF-IDF models

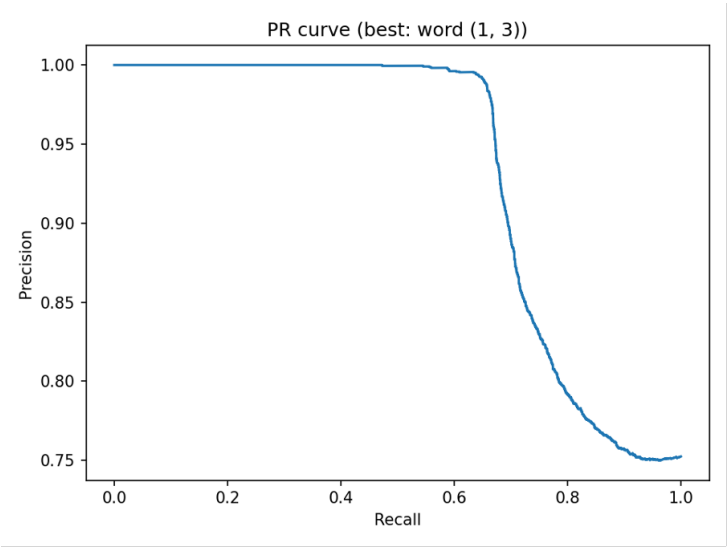


Fig.2. Precision–Recall (PR) curve for TF-IDF models

As shown in Figure 1, the ROC curves clearly demonstrate that configuration (1, 3) provides the largest area under the curve, which confirms its excellent recognition ability. Similarly, the curves of accuracy versus memorization level in

Figure 2 show that this model consistently maintains higher accuracy over a wide range of memorization levels, indicating stable ranking even with class imbalances.

Taken together, these results confirm that TF-IDF's word-based models provide a reliable and interpretable framework for detecting duplicates in Kazakh. The configuration (1, 3) with min_df ≈ {3, 5} and max_df ≈ {0.85, 0.95} provides the most balanced balance between accuracy and recall (PR-AUC ≈ 0.932, ROC-AUC ≈ 0.775). Therefore, the use of $\tau^* = 0$ in practical applications is not recommended. Instead, accuracy should be used when choosing a threshold value — for example, by optimizing Fß at $\beta < 1$ — or score calibration methods such as the Platt scale or isotonic regression should be used to increase decision stability.

For future work, it is recommended to explore hybrid function schemes that combine word- and character-level representations without the "wb" constraint, as well as integrate BM25's repeat ranking and dense embed pipelines to improve overall search and classification reliability.

At this stage, we evaluated static distributed word embeddings trained on Kazakh. We used pre-trained 300-dimensional vectors from FastText (cc.kk.300.bin) and Word2Vec (cc.kk.300.vec). Sentence vectors were obtained by either simple averaging or TF-IDF–weighted averaging of token embeddings. All vectors were L2-normalized, and pairwise similarity was computed with cosine similarity. The decision threshold τ was tuned on the validation set to maximize F1, and evaluation was performed on the test set using classification (Precision, Recall, F1) and ranking metrics (ROC-AUC, PR-AUC), together with class-wise F1 for different duplicate types.

As shown in Table 3, FastText with τ≈0.94–0.95 achieved almost perfect recall (≈1.00) but moderate precision (≈0.75), inflating F1 and generating many false positives. In contrast, Word2Vec achieved consistently near-perfect performance (F1≈0.996, ROC-AUC and PR-AUC ≈ 1.0) with both mean and TF-IDF pooling. TF-IDF weighting gave a small yet stable improvement, confirming the value of emphasizing informative tokens.

Table 3. Word-embedding baselines for Kazakh duplicate detection (zero-shot)

Metric	FastText (mean)	FastText (tfidf)	Word2Vec (mean)	Word2Vec (tfidf)
Pooling	mean	tfidf	mean	tfidf
Dim	300	300	300	300
Val_Thr	0.95	0.94	0.95	0.95
Val_P	0.735	0.738	0.9966	0.999
Val_R	0.992	0.995	0.9925	0.992
Val_F1	0.844	0.847	0.9946	0.9955
Test_P	0.747	0.749	0.9946	0.9982
Test_R	0.991	0.993	0.9941	0.9938
Test_F1	0.852	0.854	0.9943	0.996
ROC-AUC	0.844	0.844	0.9997	0.9999

PR-AUC	0.632	0.628	0.9998	0.9999
F1 [exact]	0.669	0.67	0.989	0.9963
F1 [paraphrase]	0.663	0.665	0.9893	0.9964
F1 [partial]	0.652	0.657	0.9897	0.9965
F1 [nondup]	0.885	0.886	0.9773	0.9839

FastText, with thresholds around $\tau \approx 0.94–0.95$, achieved nearly perfect recall (≈ 0.99) but only moderate precision (≈ 0.75), leading to inflated F1-scores and frequent false positives. Its class-wise results confirmed this imbalance: non-duplicates were detected reliably, while exact, paraphrase, and partial duplicates remained weak.

Word2Vec, by contrast, delivered almost flawless results. With $\tau = 0.95$, both pooling strategies reached $F1 \approx 0.996$, with ROC-AUC and PR-AUC close to 1.0, and consistently high performance across all duplicate types. TF-IDF weighting gave a small but stable improvement over mean pooling.

Thus, while FastText favors exhaustive recall, Word2Vec emerges as the more precise and robust baseline, offering a reliable foundation for Kazakh duplicate detection. Future work should confirm whether such near-perfect performance generalizes to larger, more diverse corpora.

At the final stage of the experiment, five multilingual sentence-level models were evaluated: LaBSE, multilingual-e5-base (intfloat), gte-multilingual-base (Alibaba-NLP), bge-m3 (BAAI), and snowflake-arctic-embed-l-v2.0 (Snowflake). Each model used its native pooling strategy (CLS or mean). The resulting vectors were L2-normalized, cosine similarity was used as the proximity measure, and the classification threshold was tuned on the validation set to maximize F1. On the test set we reported Precision, Recall, and F1, together with the aggregate ranking metrics ROC-AUC and PR-AUC. We also broke down F1 by example type (exact, partial, non-duplicate).

With the "validation-tuned" threshold, most models converged to similar F1 scores of about 0.670. This effect is driven by extremely high recall ($R = 1.00$) combined with moderate precision ($P \approx 0.504$), i.e., an "aggressive" duplicate decision where a low threshold labels almost all pairs as positive. Nevertheless, the models differ clearly when examined via continuous ranking metrics. By PR-AUC, BGE-M3 and Snowflake lead (≈ 0.614), followed by GTE (0.610), E5 (0.608), and LaBSE (0.594). By ROC-AUC, BGE-M3 again performs best (0.550), with Snowflake (0.545) and GTE (0.544) close behind, while LaBSE and E5 are more modest ($\approx 0.526–0.527$). These differences are especially relevant for downstream threshold calibration or for retrieval-style de-duplication.

Table 4. Results of Sentence Embedding Models on Kazakh Duplicate Detection (Zero-Shot)

Metric	LaBSE	multilingual-E5	GTE-multilingual	BGE-M3	Snowflake Arctic
Val_Thr	0.10	0.10	0.10	0.65	0.61

Val_P	0.504	0.504	0.504	0.505	0.505
Val_R	1.000	1.000	1.000	1.000	0.999
Val_F1	0.670	0.670	0.670	0.671	0.671
P (test)	0.504	0.504	0.504	0.504	0.504
R (test)	1.000	1.000	1.000	0.999	0.997
F1 (test)	0.670	0.670	0.670	0.670	0.669
ROC-AUC	0.527	0.526	0.544	0.550	0.545
PR-AUC	0.594	0.608	0.610	0.614†	0.614†
F1 [exact]	0.506	0.506	0.506	0.507	0.507
F1 [partial]	0.502	0.502	0.502	0.501	0.500
F1 [nondup]	0.000	0.000	0.000	0.002	0.005

Type-wise F1 reveals a notable pattern: scores for exact and partial duplicates are almost identical (≈ 0.506 and ≈ 0.502), indicating comparable sensitivity to exact matches and paraphrases. In contrast, the non-duplicate class remains near 0.000–0.005, showing that under the current τ^* the models effectively fail to predict negatives. Optimal thresholds cluster at low values for LaBSE, E5, and GTE ($\tau^* \approx 0.10$), but are higher for BGE-M3 and Snowflake ($\tau^* \approx 0.61$–0.65). Despite these different optima, final F1 remains similar across models, reinforcing that ranking metrics are more informative for comparing sentence-level models in a zero-shot setting.

From a practical standpoint, BGE-M3 and Snowflake are the most promising choices. Their advantages on ROC-AUC and PR-AUC suggest that, once the threshold is calibrated toward higher precision, these models have the greatest potential to improve the Precision–Recall balance. In deployment, one should not optimize F1 on validation alone; instead use criteria such as Precision@Recall $\geq r_0$ or Fβ with $\beta < 1$ to emphasize precision and avoid suppressing the non-duplicate class.

To further improve separation of negatives, it is advisable to apply hard-negative mining, light contrastive fine-tuning on paired examples from the training corpus, and hybrid architectures that combine BM25 retrieval with dense re-ranking. Overall, the results indicate that all sentence-level models form a strong recall-oriented zero-shot baseline, while BGE-M3 and Snowflake retain clear leadership on ranking quality—making them the most rational choices for threshold calibration and retrieval scenarios.

The study considers two scenarios for duplicate detection: (A) binary classification of sentence pairs and (B) retrieval. In all experiments, multilingual sentence embeddings (specifically, LaBSE and multilingual-e5-base) are used. Vectors are L2-normalized, and cosine similarity serves as the primary proximity measure.

Pair Classification (duplicate vs. non-duplicate)

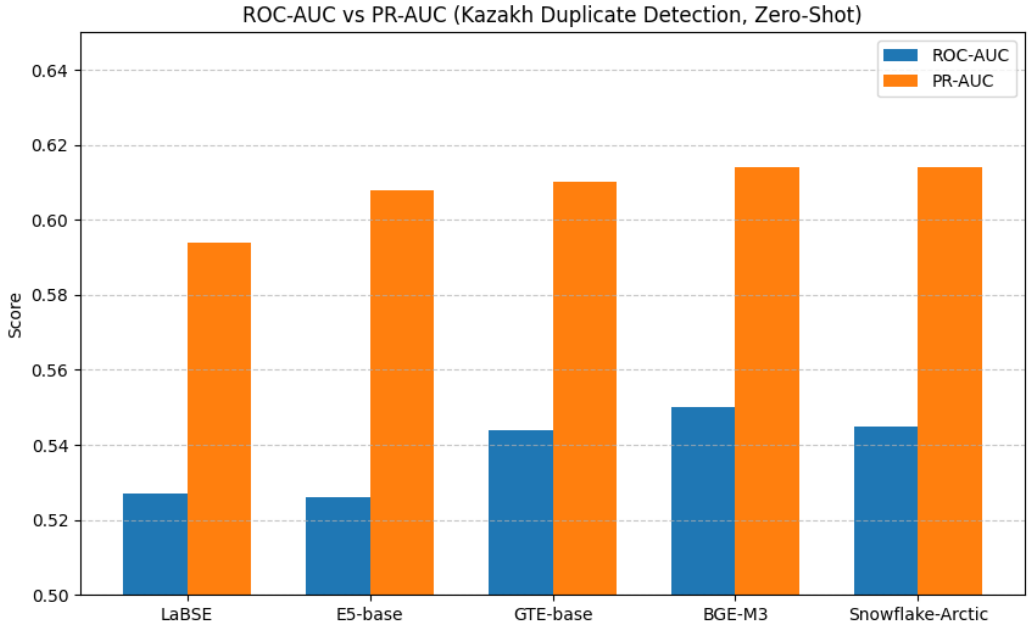


Fig.3. ROC-AUC vs PR-AUC

With thresholds tuned on validation (τ^*=0.10), the models achieved F1≈0.86, Recall=1.00, and Precision≈0.75, alongside strong ranking separability (ROC-AUC 0.76–0.77, PR-AUC ≈0.93). This confirms that while the models distinguish duplicates from non-duplicates effectively in ranking space, the chosen operating point is deliberately skewed toward maximal recall.

A robustness check with Euclidean and Manhattan distances clarified an important methodological point: because embeddings are L2-normalized, these distance measures induce rankings nearly identical to cosine. As a result, ROC-/PR-AUC values remain unchanged. However, reusing the cosine threshold grid directly in distance space yields degenerate predictions (P=R=F1=0), since the scale is mismatched. This highlights the critical need to either (i) transform distances into similarities (e.g., sim = 1 – d) or (ii) tune thresholds directly within the metric's scale.

Overall, the results emphasize the distinction between ranking ability (strong across all metrics) and cut-off calibration(sensitive to τ). In deployment, τ should be calibrated for the desired trade-off—for example, maximizing Fβ with $\beta<1$ to emphasize precision, or enforcing a Precision@Recall$\geq r_0$ constraint.

Retrieval (duplicate search)

In the retrieval setting, two pipelines were compared: a dense-only FAISS index and a hybrid BM25 followed by dense re-ranking strategy. Both LaBSE and E5 showed very similar behavior. The hybrid scheme consistently improved performance, yielding +0.1–0.2 pp gains on Recall@k and small but stable improvements in MRR and nDCG@10 compared to dense-only search. Gains were most visible in the top-10 ranking region, confirming that hybrid reranking concentrates relevant candidates more effectively.

Between models, LaBSE demonstrated a slight but consistent advantage over E5, especially on ranking-oriented metrics (MRR, nDCG@10). Importantly, Recall curves saturated after k≈10, suggesting that most relevant duplicates are captured early, and extending candidate sets beyond top-10 provides diminishing returns.

These findings support H3, showing that a hybrid BM25 combined with dense re-ranking pipeline achieves more balanced retrieval by combining lexical and semantic signals, even if the absolute improvements are modest.

Table 5. Results of binary classification of sentence pairs (duplicate vs. non-duplicate) using sentence embeddings.

Model	Sim	Val_Thr	Val_ Precision	Val_ Recall	Val_F1	Val_ F1.0	Test_Precision	Test_ Recall	Test_ F1	Test_ROC_ AUC	Test_PR_AUC
intfloat/ multilingual-e5-base	cosine	0.1	0.7524	1.0	0.8587	0.8587	0.7524	1.0	0.8587	0.7601	0.9289
intfloat/ multilingual-e5-base	euclidean	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7599	0.9289
intfloat/ multilingual-e5-base	manhattan	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7602	0.9289
sentence-transformers/ LaBSE	cosine	0.1	0.7524	1.0	0.8587	0.8587	0.7524	1.0	0.8587	0.768	0.9306
sentence-transformers/ LaBSE	euclidean	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.768	0.9306
sentence-transformers/ LaBSE	manhattan	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7694	0.931

The analysis revealed that, compared to simple mean pooling, TF-IDF–weighted aggregation provides only a marginal yet consistent improvement for the Word2Vec model (+0.001–0.002 F1; Table 3), confirming the usefulness of emphasizing informative tokens. The F1-versus-τ curve (Fig. 4) demonstrates a clear “stability window” for Word2Vec within the range $\tau \in [0.93; 0.96]$; beyond these limits, the metrics predictably drift toward precision–recall extremes. FastText, by contrast, exhibits a narrower optimum, indicating higher sensitivity to threshold calibration. Analysis of FastText false positives highlights recurring error patterns—morphological variants with minimal affix changes, named entities and toponyms occurring in similar contexts, formulaic expressions or clichés, and unattributed quotations—suggesting an overreliance on surface similarity. Finally, the metric robustness check (Fig. 5) confirms that cosine, Euclidean, and Manhattan distances

produce equivalent rankings under L2-normalization; however, directly transferring cosine thresholds to distance scales leads to degenerate predictions, underscoring the need either to convert via sim = 1 − d or to re-tune thresholds within each metric's native scale.

Table 6. Results of duplicate retrieval using sentence embeddings.

k	Recall@k	MRR	nDCG@10	Model	System
1	0.7396335583413693	0.7439812708094347	0.7396335583413693	intfloat/multilingual-e5-base	dense-only
5	0.7490838958534234	0.7439812708094347	0.745141980841072	intfloat/multilingual-e5-base	dense-only
10	0.7500482160077145	0.7439812708094347	0.7454452067553312	intfloat/multilingual-e5-base	dense-only
50	0.7515911282545805	0.7439812708094347	0.7454452067553312	intfloat/multilingual-e5-base	dense-only
1	0.740983606557377	0.745185642467465	0.740983606557377	intfloat/multilingual-e5-base	bm25→rerank
5	0.7498553519768563	0.745185642467465	0.7462413546492682	intfloat/multilingual-e5-base	bm25→rerank
10	0.7504339440694311	0.745185642467465	0.7464307724975715	intfloat/multilingual-e5-base	bm25→rerank
50	0.7523625843780135	0.745185642467465	0.7464307724975715	intfloat/multilingual-e5-base	bm25→rerank
1	0.7419479267116683	0.7451316619387559	0.7419479267116683	sentence-transformers/LaBSE	dense-only
5	0.7486981677917068	0.7451316619387559	0.7458319625713553	sentence-transformers/LaBSE	dense-only
10	0.7500482160077145	0.7451316619387559	0.7462723023088897	sentence-transformers/LaBSE	dense-only
50	0.751976856316297	0.7451316619387559	0.7462723023088897	sentence-transformers/LaBSE	dense-only
1	0.743297974927676	0.7464302433878133	0.743297974927676	sentence-transformers/LaBSE	bm25→rerank
5	0.7500482160077145	0.7464302433878133	0.747173558762129	sentence-transformers/LaBSE	bm25→rerank
10	0.7512054001928641	0.7464302433878133	0.7475570912869531	sentence-transformers/LaBSE	bm25→rerank
50	0.7523625843780135	0.7464302433878133	0.7475570912869531	sentence-transformers/LaBSE	bm25→rerank

Conclusion

This study presents the systematic comparison of TF-IDF, word, and sentence embeddings for duplicate detection in Kazakh texts. The findings reveal distinct differences in accuracy, robustness, and semantic generalization. Word2Vec with TF-IDF weighting achieved the highest and most stable performance across duplicate types, serving as a strong baseline. Sentence embeddings (notably BGE-M3 and Snowflake Arctic) excelled in capturing semantic and contextual similarities, validating their suitability for paraphrased duplicates. TF-IDF models remained competitive on exact and partial overlaps but declined on semantic cases, while FastText favored recall at the cost of precision. A BM25 combined with dense re-ranking pipeline further improved retrieval metrics, balancing lexical and semantic similarity. Overall, the results establish Word2Vec as a robust baseline and demonstrate that calibrated sentence embeddings and hybrid methods offer

superior scalability for deduplication in Kazakh and other morphologically rich, low-resource languages.

REFERENCES

Ayazbayev D., Bogdanchikov A., Orynbekova K., Varlamis I. (2023). Defining semantically close words of Kazakh language with distributed system Apache Spark // Big Data and Cognitive Computing. — 2023. — Vol. 7. — No. 4. Article 160. DOI: 10.3390/bdcc7040160.

Akhmed-Zaki D., Mansurova M., Madiyeva G., Kadyrbek N., Kyrgyzbayeva M. (2021). Development of the information system for the Kazakh language preprocessing // Cogent Engineering. — 2021. — Vol. 8. — No. 1. DOI: 10.1080/23311916.2021.1896418.

Bojanowski P., Grave É., Joulin A., Mikolov T. (2017). Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — Pp. 135–146. DOI: 10.1162/tacl_a_00051.

Bakiyev B. (2022). Method for determining the similarity of text documents for the Kazakh language, taking into account synonyms: Extension to TF-IDF // 2022 International Conference on Smart Information Systems and Technologies (SIST). — IEEE, 2022. — Pp. 1–6. DOI: 10.1109/SIST54437.2022.9945747.

Biloshchytska S., Tleubayeva A., Kuchanskyi O., Biloshchytskyi A., Andrashko Y., Toxanov S., Mukhatayev A., Sharipova S. (2025). Text similarity detection in agglutinative languages: A case study of Kazakh using hybrid n-gram and semantic models // Applied Sciences. — 2025. — Vol. 15. — No. 12. — Article 6707. DOI: 10.3390/app15126707.

Chen J., Xiao S., Zhang P., Luo K., Lian D., Liu Z. (2024). BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation // arXiv preprint. — 2024. — arXiv:2402.03216. DOI: 10.48550/arXiv.2402.03216.

Cheng L., Yang Y., Zhao K., Gao Z. (2020). Research and improvement of TF-IDF algorithm based on information theory // The 8th International Conference on Computer Engineering and Networks (CENet2018). Advances in Intelligent Systems and Computing / Eds.: Q. Liu, M. Mısır, X. Wang, W. Liu. — Springer, 2020. — Vol. 905. — Pp. 1273–1280. DOI: 10.1007/978-3-030-14680-1_67.

Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. (2020). Language-agnostic BERT sentence embedding // arXiv preprint. – 2020. — arXiv:2007.01852. DOI: 10.48550/arXiv.2007.01852.

Li P., Qiao T., Guang Y., Zhang L. (2021). A new shingling similar text detection algorithm // Advances in Simulation and Process Modelling. ISSPM 2020. Advances in Intelligent Systems and Computing / Eds.: Y. Li, Q. Zhu, F. Qiao, Z. Fan, Y. Chen. – Springer, 2021. — Vol. 1305. — Pp. 93–104. DOI: 10.1007/978-981-33-4575-1_9.

Mansurova A., Mansurova A., Nugumanova A. (2024). QA-RAG: Exploring LLM reliance on external knowledge // Big Data and Cognitive Computing. — 2024. — Vol. 8. — No. 9. — Article 115. DOI: 10.3390/bdcc8090115.

Mansurova M., Rakhimova D. (2025). Morphological parsing of Kazakh texts with deep learning approaches // Journal of Mathematics, Mechanics and Computer Science. — 2025. — Vol. 124. — No. 4. — Pp. 48–58. DOI: 10.26577/JMMCS2024-v124-i4-a4.

Mussiraliyeva S., Bolatbek M., Yeltay Z., Azanbay K. (2024). Development of an error correction algorithm for Kazakh language // Journal of Mathematics, Mechanics and Computer Science. — 2024. — Vol. 123. — No. 3. — Pp. 81–97. DOI: 10.26577/JMMCS2024-v123-i3-8.

Tleubayeva A. (2025). Kazakh Text Duplicates: a dataset for duplicate detection in Kazakh [Электронный ресурс]. — 2025. — URL: https://huggingface.co/datasets/Arailym-aitu/KazakhTextDuplicates (дата обращения: 02.09.2025).

Yu P., Merrick L., Nuti G., Campos D. (2024). Arctic-Embed 2.0: Multilingual retrieval without compromise // arXiv preprint. — 2024. — arXiv:2412.04506. DOI: 10.48550/arXiv.2412.04506.

Wang L., Zhang L., Jiang J. (2024). Duplicate question detection with deep learning in Stack Overflow // IEEE Access. – 2020. – Vol. 8. – P. 25964–25975. – DOI: 10.1109/ACCESS.2020.2971549.

Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Multilingual E5 text embeddings: a technical report // arXiv preprint. — 2024. — arXiv:2402.05672. DOI: 10.48550/arXiv.2402.05672.

Xu Z., Mo F., Huang Z., Zhang C., Yu P., Wang B., Lin J., Srikumar V. (2025). A survey of model architectures in information retrieval // arXiv preprint. — 2025. arXiv:2502.14822. DOI: 10.48550/arXiv.2502.14822.

Zhang X., Zhang Y., Long D., Xie W., Dai Z., Tang J., Lin H., Yang B., Xie P., Huang F. et al. (2024). mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval // Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. — 2024. — Pp. 1393–1412. — ACL.

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Правила оформления статьи для публикации в журнале на сайте:
https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (Казахстан, Алматы)

Подписано в печать 15.12.2025.
Формат 60x881/8. Бумага офсетная. Печать - ризограф. 9,0 п.л. Тираж 100
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание Международного университета информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59